The impact of splicing related constraints on exonic evolution

Wu, Xianming

Award date:
2016

Awarding institution:
University of Bath

The impact of splicing related constraints on exonic evolution

XianMing Wu

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

September 2015

Table of Contents

Acknowledgements

It is really very lucky that I can be a student of Laurence, and the unlucky thing is that I met him so late in my life. Under his 3-year supervision, I accept a thorough and systematic training to be an independent scientist, he is the first person who can show me that doing research is not a simple and boring job for living, rather it is a so beautiful and valuable career to the world. His personal culture, knowledge and scientific attitude merge into a calm lake with unpredictable depth; I just walk around it and try my best to drink its nutrition-enriched water everyday. I can feel that I become more "healthy" and "strong" gradually; and the most reluctant thing for me is that someday I have to leave. At the moment, the word "thank you" is too pale and weak to express my appreciation to him. I just hope that I won't be the student who will be disappointing his supervisor ☺.

I am quite grateful to University of Bath for giving me a chance for PhD study and providing me the prestigious studentship. I sincerely wish that I could do some contributions to the university in the future. I would also like to thank my friends and colleagues, especially for my beautiful English teacher Becky, for leaving me colourful memories of Bath life.

Finally, I thank my parents, wife and son (although he called me "uncle" at heathrow airport when he was 2 years old. What a horrible world, when you select to be a PhD sometimes you have to give up to be a father!) for their supporting and understanding. I feel quite ashamed that I have not fulfilled my responsibilities. I hope this thesis could be the best present for them to celebrate family reunion new year.

Contributions

Unless otherwise stated, all analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst. In Chapter II, Ana Tronholm, Eva Fernández Cáceres, Jaime M. Tovar-Corona, Lu Chen, and Araxi O. Urrutia did the analyses for "Alternative Splicing Event Calculation" part.

Summary

Regulation of pre-mRNA splicing is a key process for most if not all eukaryotes. The process can, in the abstract, be considered as a series of *trans*-acting factors that interact with *cis*-motifs in the RNA to enable the removal of introns and joining of exons. As the *cis* factors need not only be the splice sites themselves, but also motifs in the exons, the splicing process has the potential to impose selective constraint on exonic sequence in addition to the normal selection on the amino acid content of the protein. To understand this more clearly, in this thesis, I mainly focus on a type of important and widely investigated *cis*-motifs, exonic splicing enhancers (ESEs), which bind with SR proteins to re-enforce the splice sites and so ensure splicing correctly. First, I explore splice-related *cis*-motif usage of the *Ectocarpus* genome, which is a species phylogenetically very distant from vertebrates but, like vertebrates in having abundant large introns. A deep phylogenetic conservation of exonic splice-related constraints is observed (Chapter II). Then I extend the analysis across taxa in a phylogenetically explicit framework. In this section stronger selection on exon end synonymous sites can be detected within humans when the exons are flanked by larger introns. Additionally I report evidence that reduced N_e might lead to larger introns and weakened splice sites. Thus I suggest an unusual circumstance in which selection (for *cis*-motifs to control error-prone splicing) might be stronger when population sizes are smaller; this is unexpected and would be a necessary complement to nearly-neutral theory (Chapter III). Third, I ask whether what we know about biases in the usage of ESEs and splicing control elements allows us to understand where in human genes pathogenic mutations tend to occur (Chapter IV). By examining the relationship between determinants of the usage of splice-associated *cis*-motifs and the distribution of human pathogenic SNPs, I found certain exons are vulnerable to splice disruption owing to low ESE density and a "fragile" exon model we proposed could describe and explain this phenomenon (Chapter IV). Finally I perform preliminary analysis, with a view to biotechnological optimization of transgenes, to address whether there might be such a thing as a tissue specific ESE. To this end I examine ESE usage in tissue specific genes. I find some preliminary evidence for tissue specific biased usage of certain ESEs.

Abbreviations

- A - adenine
- C – cytosine
- CAI - Codon Adaptation Index
- CDS - coding DNA sequence
- ESE - Exonic Splicing Enhancer
- ESS - Exonic Splicing Silencer
- EST - Expressed Sequence Tag
- G - guanine
- Ks - synonymous substitution rates
- MCMC - Markov Chain Monte Carlo
- mRNA - messenger RNA
- N_e - effective population size
- NMD - Nonsense Mediated Decay
- RESCUE - Relative Enhancer and Silencer Classification by Unanimous Enrichment
- SNP - Single Nucleotide Polymorphism
- SR proteins - the serine/arginine-rich proteins
- T - thymine
- tRNA - transfer RNA

Chapter I. Introduction

In eukaryotes, pre-mRNA splicing, the removal of introns and joining of exons, is commonplace in many taxa. In humans for example probably around 97% of protein coding genes have at least two exons (Grzybowska 2012). Normally, splicing process occurs in spliceosome, a large ribonucleoprotein (RNP) machine composed of five small nuclear ribonucleoprotein (snRNP) complexes (U1, U2, U4/U6, and U5) and functions by dynamic assembly and disassembly cycle (Lee and Rio 2015)(Fig. 1). Control of splicing is often central to phenotype specification. Splicing patterns of genes in the sex determination pathway, for example, define the sex of fruit flies (Grzybowska 2012). But how are splice sites recognized and how, if at all, does their mode of recognition impact on gene evolution?

Previously, exon-intron boundaries were thought to be recognized by simple patterns: AG dinucleotide splice acceptor, GT dinucleotide splice donor and an intronic branch site (Kent and Zahler 2000; Black 2003) (Fig. 2). In this model there is no exonic specification of splice sites beyond the one of two base pairs at the splice site. However, in many taxa these features are not sufficient for correct gene splicing and *cis*-motifs at exon ends (within ~70-100bp of splice site) provide reinforcement and definition of the flanking splice sites, especially if such sites are "weak" (Berget 1995; Graveley 2000; Fairbrother et al. 2002; Dewey, Rogozin, and Koonin 2006; Plass et al. 2008; Cáceres and Hurst 2013). It is estimated that in the human genome, approximately 50% of the information defining splice sites is in the *cis* motifs (Lim and Burge 2001).

Two important exonic splicing control elements, serving as enhancers (Exonic Splicing Enhancers, ESEs) (Blencowe 2000) and silencers (Exonic Splicing Silencers, ESSs) (Amendt, Si, and Stoltzfus 1995; Kan and Green 1999), are considered. They reside within exon ends and affect (promote or inhibit) the exact identification of splice sites by interacting with certain protein regulators (SR proteins and hnRNP) (Zheng et al. 2000; Rowen et al. 2002) (Fig. 3). Exon end ESEs generally are under purifying selection (Parmley and Hurst 2007). For example, from substitutional data it is estimated that about 4-5% of synonymous mutations are under selection in humans because they disrupt ESEs (Cáceres and Hurst 2013). Similarly, new point mutations (SNPs) at exon ends are likely to be eliminated by purifying selection if they disrupt known motifs (Majewski and Ott 2002;

Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley, Chamary, and Hurst 2006; Parmley and Hurst 2007; Cáceres and Hurst 2013; Wu and Hurst 2015). The involvement of *cis*-motifs also distorts codon usage at exonic ends in a manner that is predictable from the nucleotide content of splice-associated exonic motifs (Parmley and Hurst 2007; Cáceres and Hurst 2013). However, there is little evidence that selection acts upon ESS motifs (Chamary, Parmley, and Hurst 2006; Parmley and Hurst 2007; Parmley et al. 2007). Thus, in this thesis, I concentrate my attention on ESEs.

Previously researchers have identified candidate ESE sequences by laborious experimental methods, for instance, by site-directed mutagenesis of minigene constructs and by protocols based on SELEX (Systematic Evolution of Ligands by Exponential enrichment). In so doing they have been able to identify several sequences with enhancer activity from a pool of random sequences (Tian and Kole 1995; Coulter, Landree, and Cooper 1997; Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999; Liu et al. 2000). Bioinformatics methodology revolutionized the hunt for candidate ESEs by, for example, enabling understanding the disruption of ESEs through the analysis of disease alleles (Cartegni, Chew, and Krainer 2002) and by identification of *k*-mer sequences that are especially abundant near exon ends (Lim et al. 2011). The most commonly employed current methodology starts with *in silico* methods that identify *k*-mers (typically 6-mers) enriched in multiple orthogonal dimensions, followed by experimental validation (RESCUE-ESE) (Fairbrother et al. 2002). Fairbrother et al, for example searched for hexamers enriched in exons compared with introns and in exons associated with weak splice sites versus strong splice sites (Fairbrother et al. 2002; Fairbrother et al. 2004b). The candidate list of hexamers enriched on both dimensions were then subject to experimental testing.

These studies of ESEs have led to a broad consensus as to the properties of ESEs which include: 1) ESEs generally locate in the vicinity of splice sites (Berget 1995; Fairbrother et al. 2004a) and appear to be functional up to around 70 nucleotides from an exon end (Fairbrother et al. 2004a); 2) the main known function of ESEs is enhancing splicing and reinforcing recognition of the correct neighbouring splice site (Blencowe 2000; Graveley 2000), although in some contexts enhancers can behave as silencers for reasons still unclear (Ke et al. 2011). 3) ESEs exert influence at the immature RNA level by binding with serine and arginine-rich (SR) proteins (Graveley 2000). Generally, these SR proteins are 50-300 amino acids in length and composed of two domains, the RNA recognition motif (RRM) region and the splicing machinery binding domain, while many unexpected functions have been found recently, such as playing roles on RNA transcription, export,

translation, and decay (Howard and Sanford 2015). 4) ESEs work by some form of mass action as ESE density is typically rather high (using the RESCUE-ESE set of hexamers on average 30-40% of sequence near exon ends matches known ESEs (Parmley, Chamary, and Hurst 2006)). 5) ESEs may be more prevalent in species with large introns (Warnecke, Parmley, and Hurst 2008) and at a higher density at exon ends in proximity to longer introns (Dewey, Rogozin, and Koonin 2006; Cáceres and Hurst 2013); 6) ESE sequence is likely to be under purifying selection (see above). 7) ESEs tend to be purine enriched (Tanaka, Watakabe, and Shimura 1994; Fairbrother et al. 2004a; Parmley et al. 2007; Cáceres and Hurst 2013). In the RESCUE ESE data set for example 50% of nucleotides are A, 25% G giving a 75% purine loading.

A major concern is needed here as most of the data pertinent to the impact and properties of ESEs comes from mammals, one of the few taxa in which ESEs have been experimentally confirmed. How then can we explore the impact of ESEs in non-mammalian taxa? Importantly, that ESEs have a highly skewed nucleotide usage with a strong preference for A and purines more generally, enables us to measure usage of *cis* splice motifs based on nucleotide content at exon ends. Specifically codon usage at exon ends is well predicted by underlying nucleotide biases in ESEs (Parmley and Hurst 2007; Cáceres and Hurst 2013). This property is useful as it means that we can attempt to understand the extent of *cis*-motif usage in taxa without experimentally defined motifs by examining trends in codon and amino acid usage at exon ends. Indeed, when between-species variation in exonic *cis*-motif usage is considered in this thesis, we presume that the frequency of distorted codon or amino acid usage in the vicinity of exon junctions is a fair measure.

The trend in usage of each codon and amino acid is investigated as a function of the distance from the exon–intron boundary up to a distance of 34 codons (to accord with an earlier analysis (Warnecke, Parmley, and Hurst 2008)). That is to say, for any given species, I consider the usage of any given codon in well-annotated exons, as a function of the distance from the boundary. A codon that is associated with ESEs should be one increasing in usage as one approaches the boundary. By contrast a codon not commensurate with ESE usage should be one avoided near exon ends. I consider typically the spearman correlation (rho) between relative codon usage and distance. A positive rho value indicates a codon that is avoided, a negative value one that is preferred near exon ends, likely one involved in ESE specification. To ask about the extent to which ESEs are used in a species I employ the proportion of codons or amino acids that show significant

trends in their usage as a function of the distance from an exon-intron junction. These metrics have been shown, by comparison with reference ESE sets, to correspond well with ESE motif usage (Parmley and Hurst 2007; Parmley et al. 2007; Warnecke, Parmley, and Hurst 2008; Cáceres and Hurst 2013).

Given the unusual nucleotide content, their high density and their exposure to purifying selection, ESE and *cis*-motif selection more generally has the potential to profoundly influence coding sequence evolution. Here I extend the above findings to address further issues in the relationship between ESE functioning and fitness. In particular I address four questions. First is it generally the case that species with larger introns use ESEs more than other taxa? I examine the case of *Ectocarpus* in an in depth analysis (Chapter II) and then extend the analysis across taxa in a phylogenetically explicit framework (Chapter III). Second, given that ESE usage is higher when introns are large, and that introns are large when population sizes are small, does it follow that selection for accurate splicing is stronger when populations are small (in contradiction of the more common assumption that selection is weakest in small populations) (Chapter III)? Third, I ask whether what we know about biases in the usage of ESEs and splicing control elements allows us to understand where in human genes pathogenic mutations occur (Chapter IV). Finally I ask whether there might be such things as relative tissue specific ESEs (Chapter V).

1.1 Do species with big introns use ESEs more?

The possible connection with intron size mentioned above is central to many aspects of my thesis. Experimental insertion of sequence into introns tends to reduce the rate at which the intron is spliced correctly (Klinz and Gallwitz 1985; Luehrsen and Walbot 1992; Fox-Walsh et al. 2005; Sironen et al. 2006). Likewise, exons flanked by large introns tend to be phylogenetically lost (possibly owing to missplicing) (Kandul and Noor 2009). Splicing is also considered more noisy when introns are large (Bell et al. 1998; Fox-Walsh et al. 2005). All this data has led to the notion that large introns are hard to splice accurately and so need reinforcement. This reinforcement comes in the form of exonic splice enhancers, so explaining the possible coupling with intron size within and between genomes. However, the only species with large introns examined to date have been mammals. To understand ESEs usage in phylogenetic perspective, especially for intron-rich species besides mammals, I start by examining the patterns of codon and amino acid usage in the vicinity of exon-intron junctions in the brown algae *Ectocarpus siliculosus*.

This species is unusual in that the genome is well sequenced and annotated (Cock et al. 2010; Cock et al. 2012)), it is a species with abundant large introns, known SR proteins and classical splice sites (Cock et al. 2010) (Chapter II). In this chapter I thus ask how common splice-related skews at exonic ends might be in *Ectocarpus* and how they compare with those seen in humans. This provides the first dissection of *cis*-motif usage in a species distant from humans but comparable in genome anatomy.

Aside from simply assessing the commonality of trends, I also attempt to define ESEs for this species. The most commonly employed set of human ESEs are those derived by the RESCUE-ESE methodology (Fairbrother et al. 2004b), which looks for motifs enriched along multiple orthogonal axes. I attempt a similar method to assemble a set of *Ectocarpus* putative ESEs (see method of Chapter II). I then compare such motifs to human motifs to see if there is any resemblance.

As the 3-mer in frame (codons) usage trends in mammals relate to the nucleotide content of ESEs (Parmley and Hurst 2007), I also explore, in *Ectocarpus*, whether exon end codon usage bias relates to distribution of ESEs and whether putative hexameric ESEs have significant difference with deep phylogenetic SR protein binding motifs. Furthermore, given usage trends, in mammals, at the 5' and 3' ends of exons appear to be largely symmetrical (if a codon or amino acid is highly preferred at the 5' end of exons, it is similarly highly preferred at the 3' end) (Warnecke, Parmley, and Hurst 2008; Lim et al. 2011), but antisymmetry trends were observed in *Caenorhabditis* worms (Warnecke, Parmley, and Hurst 2008), I ask whether symmetry is seen in *Ectocarpus*. I find that *Ectocarpus* is very rich in ESEs, these ESEs tend to be symmetrical and resemble those seen in humans. This suggests very deep conservation of ESEs and that the trend for intron-rich species to have many ESEs to be found outside of mammals.

Are splicing optimal and translationally optimal codons mutually exclusive?

When considering codon usage it is often supposed that one is considering translational selection, wherein tRNA usage predicts codon usage, most especially in the most highly expressed genes (Duret 2002; Sharp et al. 2005). The codon matching the most abundant isoacceptor tRNA is then termed the "optimal" codon. What is the relationship between optimal codons and the distortions of codon usage at exon ends associated with ESEs? Prior evidence from *Drosophila* has suggested that splice optimal and translationally optimal codons are mutually exclusive (Warnecke and Hurst 2007). This might make sense: As codon usage trends at exon ends is a feature of usage of splicing related *cis*-motifs,

"splicing optimal" codons might be different with "translationally optimal" ones, otherwise SR proteins would have difficulty recognizing exclusively exonic ends in highly expressed genes. No attempt has been made to consider the generalizability of this result, not least because taxa with much ESE usage like humans for the most part do not have translationally optimal codons. In *Ectocarpus*, by considering trends with respect to expression level I found some codons to be "translationally optimal" (Chapter II). *Ectocarpus* thus presents a first new species with abundant ESE usage and translational selection. Here then I also make use of *Ectocarpus* to test the hypothesis that "translationally optimal" and "splicing optimal" codons are mutually exclusive.

1.2 Is selection stronger when 'N_e' is low?

The nearly-neutral theory and its possible complement

The efficacy of selection depends not only on the extent to which a new allele affects fitness, but also on the size of the population into which this mutation is introduced. The key variable is the effective population size (N_e) that (Wright 1931) defined as "the number of breeding individuals in an idealised population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration". The nearly-neutral theory predicts that a mutation will be 'effectively neutral' if its selective disadvantage (s) is negligible compared with N_e (more precisely, if $s << 1/(2\ N_e)$ for a diploid population) (Ohta 1973; Ohta 1992; Ohta 1996). According to this classical theory, there would be relatively low levels of selective constraint in species with low N_e (Ohta 1973; Ohta 1992; Ohta 1996). Then, as N_e goes down, it becomes harder to eliminate weakly deleterious insertion mutations in genomes by purifying selection. This could help us to understand why there is more noncoding DNA in human genome than in, for example, the yeast genome (Lynch and Conery 2003).

However, this genome decay process may lead to new problems, such as increased mistranscription, mistranslation, missplicing, incorrect protein folding, incorrect phosphorylation, incorrect subcellular localization *etc*. (Lynch 2007). These errors might necessitate the evolution of protection mechanisms. This means that, as N_e reduces, unexpectedly stronger selection may be imposed on error mitigation. We observe that prior data suggests that a) introns may be longer when N_e is low (Lynch and Conery 2003) and b) that longer introns are associated with a greater need to control splicing, intron size being experimentally shown to affect splice accuracy (Klinz and Gallwitz 1985; Luehrsen and Walbot 1992; Bell et al. 1998; Fox-Walsh et al. 2005; Sironen et al. 2006; Kandul and

Noor 2009). Thus it might be reasonable to ask whether selection on ESEs is stronger when introns are large and populations small. To test this possibility, I focus on selection constraints related to splicing error in both intragenomic and intergenomic comparisons (Chapter III).

Are more synonymous sites under selection near long introns?

In our model, introns undergo expansion due to multiple and gradual small insertions, each being unable to be resisted by purifying selection if the population is small. This causes increased selection on modifiers of splicing in a ratchet-like process (c.f. Frank 2007). The selection to reduce splice error rates we suggest will be manifested, in part, as a higher density of exonic *cis*-modifiers of splicing in proximity to exons with large introns and in species with larger introns on the average.

We can employ data from primates to examine the intra-specific prediction. As mentioned above, these splicing control *cis*-motifs are enriched towards the ends of exons (Fairbrother et al. 2004a), which in turn cause selective constraint at synonymous sites (Carlini and Genut 2006; Parmley, Chamary, and Hurst 2006) and a highly skewed nucleotide usage (Tanaka, Watakabe, and Shimura 1994; Fairbrother et al. 2004a; Parmley et al. 2007). We thus estimate the proportion of sites at exon ends that are in ESE and under purifying selection as a function of intron size. Determining significance is non-trivial owing to the skewed nucleotide content of ESEs. To mitigate this issue, we do simulation of randomized pseudoESE sets that are the same size and drawn from the same underlying nucleotide content as the true ESE sets.

Do genomes with large introns exonize splicing information?

To examine the inter-specific predictions we start by re-evaluating the connection between $N_e.\mu$ and intron size, this being central to our model. From phylogenetically uncontrolled correlation based analysis Lynch and Conery (2003) noted that across a wide span of species, as $N_e.\mu$ declines introns tend to get larger and more common (higher density). $N_e.\mu$ note is the product of effective population size (N_e) and the mutation rate (μ), the single statistic being estimated from population heterozygozity data. The trend in intron size Lynch and Conery attribute to weakening selection as N_e declines, i.e. species with low N_e are less able to eliminate, via purifying selection, weakly deleterious insertion mutations when they occur in introns (and intergenic sequence). This study has, however, been criticised for failing to allow for phylogenic non-independence between data points (Whitney and Garland 2010). Indeed, it was argued that the key result is not robust to

proper phylogenetic control (Whitney and Garland 2010). As this $N_e.\mu$ intron size/number correlation is a central tenet of the nearly-neutral interpretation of genome anatomy, we return to this issue employing a phylogenetically controlled mode of analysis and more up to date estimates of $N_e.\mu$, employing both more data and multiple modes of estimation. We show that with these updated estimates, in a phylogenetically controlled framework, $N_e.\mu$ does indeed predict intron dimensions as Lynch and Conery (2003) postulated. We also show, however, that Whitney and Garland had an important objection, as we do not robustly recover this result using the original Lynch and Conery estimates of $N_e.\mu$.

Given this result, I then move to testing whether ESE usage, intron size and intron density covary and whether the trends are predicted by N_e (or rather $N_e.\mu$). Unexpectedly I discover that, in addition to intron size, intron density is a very strong predictor of ESE usage and in turn suggest a new synthetic model for which genes/exons and which species might be especially enriched for ESEs.

1.3 Does ESE usage predict the intragene location of pathogenic mutations?

The above analyses lead to a synthetic view of which genes and exons have great problems splicing correctly and which in turn require more ESEs to reinforce the splicing process. Do these predictors in turn enable us to understand where in human genes disease-causing mutations occur? Given, for example, that ESEs function at exon ends, are disease-causing mutations more common at exon ends? If so how much more common are they and with such information can we estimate how many disease causing mutations mediate effects via splicing?

In addition, we can ask whether within genes disease-associated mutations are particularly associated with ESE rich exons and ESE poor ones or does it make no difference? What then more generally is the relationship between determinants of ESEs usage and the distribution of pathogenic SNPs. This question is potentially of importance if we wish to infer which SNPs might be pathogenic and which not. Given the recognition that even synonymous SNPs cause disease, often by disrupting splicing (Faustino and Cooper 2003; Chamary, Parmley, and Hurst 2006), converting such insights into improved detection of pathogenic SNPs seems like a pressing immediate concern. In order to carry this analysis, I consider five correlates of ESEs usages (Chapter IV):

1) *relative position in exons (flank versus core)*

ESEs are enriched at exon ends and are under selection in these domains. In flanking regions of exons, both substitution rates and polymorphism rates are lower than that in core regions (Majewski and Ott 2002; Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley, Chamary, and Hurst 2006; Parmley et al. 2007; Cáceres and Hurst 2013; Wu and Hurst 2015). This is partly explained by richness of functional ESEs in exonic flanks (Nelson and Green 1988; Lavigueur et al. 1993; Graveley, Hertel, and Maniatis 1998; Fairbrother et al. 2004a; Carlini and Genut 2006; Parmley, Chamary, and Hurst 2006; Parmley et al. 2007; Cáceres and Hurst 2013).

2) *relative position in genes (5' versus 3')*

In a gene, exons at the 5' end have more downstream splice sites than ones downstream and hence have a larger number of potential decoy splice sites. This logic we speculated could explain why exons in 5' positions have higher ESE density than that those in 3' positions. This is supported by comparison between second exons and last but one exons within the same gene. For example, it was observed that there is a 2 fold greater ESE density in the former (Wu and Hurst 2015).

3) *flanking intron size*

As mentioned above, species with more abundant large introns tend to have more ESEs and exons flanked by larger introns are harder to splice. Given this it is easy to understand why ESE density, in the human genome, might be higher in the exons flanked by larger introns (Dewey, Rogozin, and Koonin 2006; Cáceres and Hurst 2013; Wu and Hurst 2015).

4) *usage of splice sites (AGgt versus non AGgt)*

I found that usage of tetranucleotide splice sites "AGgt" and "agGT" (the two nucleotides in upper case come from exons and those in lower case come from introns) correlates well and positively with usage of ESEs across species (Wu and Hurst 2015). This accords with prior notions that splice site strength and ESE density coevolve (Fairbrother et al. 2002; Dewey, Rogozin, and Koonin 2006).

5) *coding phase of splice site (zero versus non-zero)*

Three possible phases (0, 1, and 2), which show the status of codons at the splice sites (e.g. phase zero exons being those cut between whole codons), are found in unequal proportions in most genomes (Fedorov et al. 1992; Long, Rosenberg, and Gilbert 1995; Ruvinsky et al. 2005). In the human genome, a significant relation between coding phases and specific splice site ("AGgt" and "agGT") usage suggests that coding phase could be a predictor of ESE density. This is also confirmed by that proportion of phase zero splice site correlates well and phylogenetically with metric of *cis*-motifs usage and intron metrics (mean intron size and intron density).

I discover that genes with many exons tend to more commonly be disease-causing genes (even controlling for CDS length) and that pathogenic mutations are greatly enriched at exon ends. Both of these observations suggest that splice disruption is a key mode of pathogenesis. Indeed I estimate that 20-45% of disease associated SNPs disrupt splicing. Importantly, for the four other predictors I find that disease-associated mutations tend to be associated with exons predicted to have low ESE density. I confirm this by showing a correlation between SNP density and ESE density. Given this I suggest the concept of the "fragile" exon, one more easily disrupted by single exon end splice disrupting mutations.

1.4 Splicing related constraints on human tissue-specific genes

For the most part my thesis has treated all ESEs as equivalent. However some may, for example, bind one SR protein better than another. This is almost certainly the case (Liu, Zhang, and Krainer 1998; Schaal and Maniatis 1999). If so, and if SR usage varies by tissue, then we might expect that tissue specific ESEs might exist. In my final chapter I define a set of human genes that appear to be expressed in only one tissue. I then employ this set to ask whether, within a set of low false positive ESE motifs, some motifs are especially abundant at exon ends in genes expressed in only certain tissues.

If such ESEs can be found they are potentially important for transgene engineering. For example, in gene therapy it is desirable for a gene to be correctly expressed often in just one tissue. Usage of highly tissue specific promoters can go some way to achieve this. But tissue specific expression is pointless if splicing cannot be done correctly. Often transgenes are synthesized lacking all but the first intron. It may thus be necessary to bolster the first exon with tissue specific splicing information, while removing such information (e.g via modification of synonymous sites) from the areas in the gene that were in proximity to exon junctions but no near intronic sequence in the transgene. This chapter provides a first analysis of whether ESEs might have tissue specific functionality.

Figures

Fig. 1 The spliceosome assembly and disassembly cycle

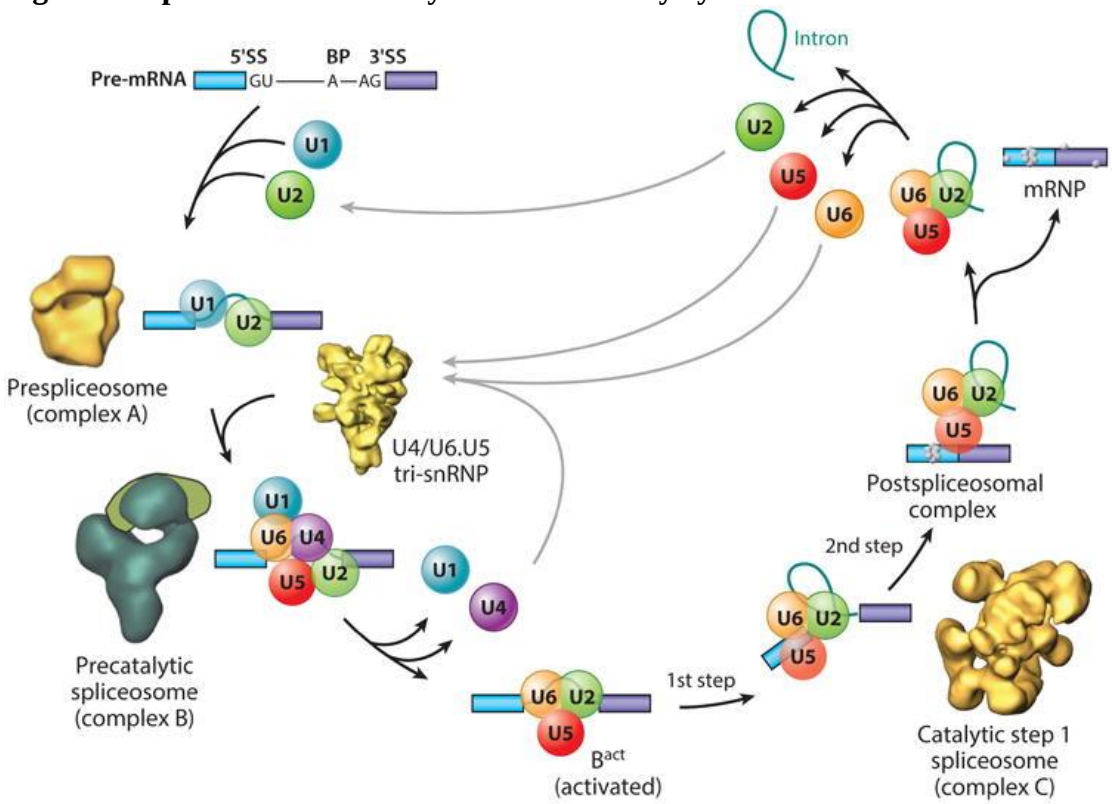


Fig. 2 Intronic terminal dinucleotides (5' GU and 3' AG) and an branch site represent important signals for recognition of exon-intron boundaries

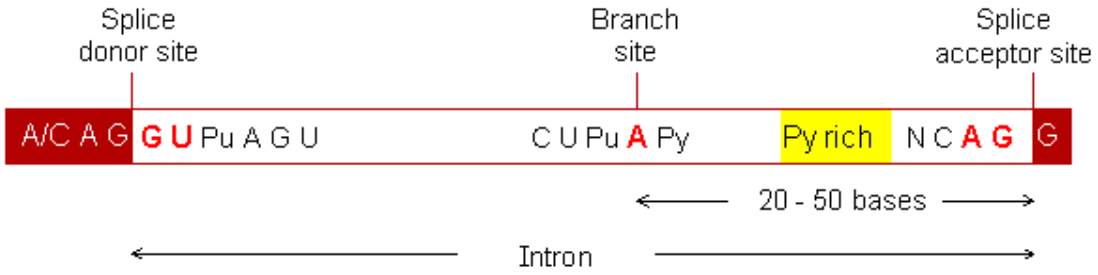
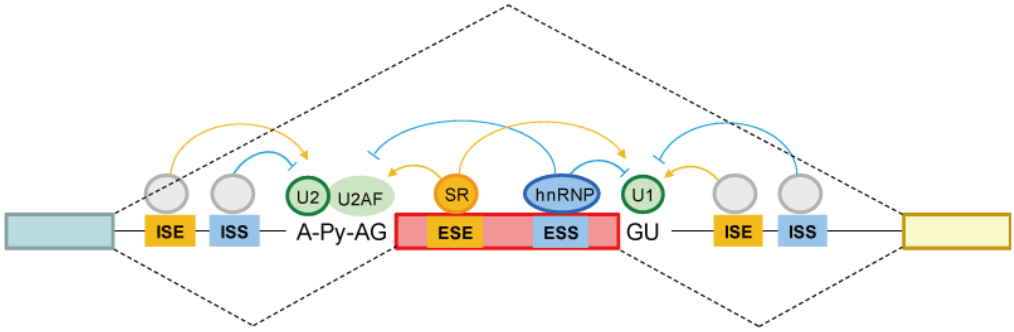


Fig. 3 Two important exonic splicing control elements (ESE and ESS) reside within exon ends and affect (promote or inhibit) the exact identification of splice sites by interacting with certain protein regulators (SR proteins and hnRNP)



References

Amendt BA, Si ZH, Stoltzfus CM 1995. Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. *Mol Cell Biol* 15: 6480.

Bell MV, Cowper AE, Lefranc MP, Bell JI, Screaton GR 1998. Influence of intron length on alternative splicing of CD44. *Mol Cell Biol* 18: 5930-5941.

Berget SM 1995. Exon recognition in vertebrate splicing. *J Biol Chem* 270: 2411-2414.

Black DL 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-336.

Blencowe BJ 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 25: 106-110.

Cáceres EF, Hurst LD 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14: R143.

Carlini DB, Genut JE 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol* 62: 89-98.

Cartegni L, Chew SL, Krainer AR 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3: 285-298.

Chamary JV, Parmley JL, Hurst LD 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7: 98-108.

Cock JM, Sterck L, Ahmed S, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Arun A, Aury JM, Badger JH, et al. 2012. The Ectocarpus Genome and Brown Algal Genomics The Ectocarpus Genome Consortium. In: Piganeau G, editor. Genomic Insights into the Biology of Algae. Amsterdam: Elsevier Ltd. p. 141-184.

Cock JM, Sterck L, Rouze P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH, et al. 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465: 617-621.

Coulter LR, Landree MA, Cooper TA 1997. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol* 17: 2143-2150.

Dewey CN, Rogozin IB, Koonin EV 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7: 311.

Duret L 2002. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics and Development* 12: 640-649.

Fairbrother WG, Holste D, Burge CB, Sharp PA 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* 2: E268.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007-1013.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* 32: W187-190.

Faustino NA, Cooper TA 2003. Pre-mRNA splicing and human disease. *Genes Dev* 17: 419-437.

Fedorov A, Suboch G, Bujakov M, Fedorova L 1992. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res* 20: 2553-2557.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* 102: 16176-16181.

Frank SA 2007. Maladaptation and the paradox of robustness in evolution. *PLoS One* 2: e1021.

Graveley BR 2000. Sorting out the complexity of SR protein functions. *RNA* 6: 1197-1211.

Graveley BR, Hertel KJ, Maniatis T 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J* 17: 6747-6756.

Grzybowska EA 2012. Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem Biophys Res Commun* 424: 1-6.

Howard JM, Sanford JR 2015. The RNAissance family: SR proteins as multifaceted regulators of gene expression. *Wiley Interdiscip Rev RNA* 6: 93-110.

Kan JL, Green MR 1999. Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. *Genes Dev* 13: 462-471.

Kandul NP, Noor MA 2009. Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC Genet* 10: 67.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 21: 1360-1374.

Kent WJ, Zahler AM 2000. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res* 10: 1115-1125.

Klinz FJ, Gallwitz D 1985. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 13: 3791-3804.

Lavigueur A, La Branche H, Kornblihtt AR, Chabot B 1993. A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. *Genes Dev* 7: 2405-2417.

Lee Y, Rio DC 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem* 84: 291-323.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A* 108: 11093-11098.

Lim LP, Burge CB 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* 98: 11193-11198.

Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR 2000. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* 20: 1063-1071.

Liu HX, Zhang M, Krainer AR 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12: 1998-2012.

Long M, Rosenberg C, Gilbert W 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc Natl Acad Sci U S A* 92: 12495-12499.

Luehrsen KR, Walbot V 1992. Insertion of non-intron sequence into maize introns interferes with splicing. *Nucleic acids research* 20: 5181-5187.

Lynch M. 2007. The origins of genome architecture. Sunderland, Mass.: Sinauer Associates ; Basingstoke : Palgrave [distributor].

Lynch M, Conery JS 2003. The origins of genome complexity. *Science* 302: 1401-1404.

Majewski J, Ott J 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res* 12: 1827-1836.

Nelson KK, Green MR 1988. Splice site selection and ribonucleoprotein complex assembly during in vitro pre-mRNA splicing. *Genes Dev* 2: 319-329.

Ohta 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst* 23: 263–286.

Ohta T 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96-98.

Ohta T 1996. The current significance and standing of neutral and neutral theories. *Bioessays* 18: 673-677; discussion 683.

Parmley JL, Chamary JV, Hurst LD 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* 23: 301-309.

Parmley JL, Hurst LD 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol* 24: 1600-1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* 5: e14.

Plass M, Agirre E, Reyes D, Camara F, Eyras E 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet* 24: 590-594.

Rowen L, Young J, Birditt B, Kaur A, Madan A, Philipps DL, Qin S, Minx P, Wilson RK, Hood L, et al. 2002. Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* 79: 587-597.

Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD 2005. Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol* 60: 99-104.

Schaal TD, Maniatis T 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol* 19: 1705-1719.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucl. Acids Res.* 33: 1141-1153.

Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J 2006. An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A* 103: 5006-5011.

Tanaka K, Watakabe A, Shimura Y 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol Cell Biol* 14: 1347-1354.

Tian H, Kole R 1995. Selection of novel exon recognition elements from a pool of random sequences. *Mol Cell Biol* 15: 6291-6298.

Warnecke T, Hurst LD 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Molecular Biology and Evolution* 24: 2755-2762.

Warnecke T, Parmley JL, Hurst LD 2008. Finding exonic islands in a sea of non-coding sequence:

splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol* 9: R29.

Whitney KD, Garland T, Jr. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet* 6.

Wright S 1931. Evolution in Mendelian Populations. *Genetics* 16: 97-159.

Wu X, Hurst LD 2015. Why Selection Might Be Stronger When Populations Are Small: Intron Size and Density Predict within and between-Species Usage of Exonic Splice Associated cis-Motifs. *Mol Biol Evol* 32: 1847-1861.

Zheng ZM, Quintero J, Reid ES, Gocke C, Baker CC 2000. Optimization of a weak 3' splice site counteracts the function of a bovine papillomavirus type 1 exonic splicing suppressor in vitro and in vivo. *J Virol* 74: 5902-5910.

Chapter II.

Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga *Ectocarpus* are common and resemble those seen in humans

Published manuscript:

Wu X, Tronholm A, Fernández Cáceres E, Tovar-Corona JM, Chen L, Urrutia AO, Hurst LD. Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga *Ectocarpus* are common and resemble those seen in humans. ***Genome Biol Evol***. 2013 Jul 30.

Contributions

In Chapter II, Ana Tronholm, Eva Fernández Cáceres, Jaime M. Tovar-Corona, Lu Chen, and Araxi O. Urrutia did the analyses for "Alternative Splicing Event Calculation" part. Other analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst.

Evidence for Deep Phylogenetic Conservation of Exonic Splice-Related Constraints: Splice-Related Skews at Exonic Ends in the Brown Alga *Ectocarpus* Are Common and Resemble Those Seen in Humans

XianMing Wu[1], Ana Tronholm[1,3], Eva Fernández Cáceres[1], Jaime M. Tovar-Corona[1], Lu Chen[2], Araxi O. Urrutia[1], and Laurence D. Hurst[1,*]

[1]Department of Biology and Biochemistry, University of Bath, Somerset, United Kingdom

[2]Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, United Kingdom

[3]Present address: Department of Biological Sciences, University of Alabama, Mary Harmon Bryant Hall, Tuscaloosa, AL

*Corresponding author: E-mail: l.d.hurst@bath.ac.uk.

Accepted: July 25, 2013

Abstract

The control of RNA splicing is often modulated by exonic motifs near splice sites. Chief among these are exonic splice enhancers (ESEs). Well-described ESEs in mammals are purine rich and cause predictable skews in codon and amino acid usage toward exonic ends. Looking across species, those with relatively abundant intronic sequence are those with the more profound end of exon skews, indicative of exonization of splice site recognition. To date, the only intron-rich species that have been analyzed are mammals, precluding any conclusions about the likely ancestral condition. Here, we examine the patterns of codon and amino acid usage in the vicinity of exon–intron junctions in the brown alga *Ectocarpus siliculosus*, a species with abundant large introns, known SR proteins, and classical splice sites. We find that amino acids and codons preferred/avoided at both 3′ and 5′ ends in *Ectocarpus*, of which there are many, tend, on average, to also be preferred/avoided at the same exon ends in humans. Moreover, the preferences observed at the 5′ ends of exons are largely the same as those at the 3′ ends, a symmetry trend only previously observed in animals. We predict putative hexameric ESEs in *Ectocarpus* and show that these are purine rich and that there are many more of these identified as functional ESEs in humans than expected by chance. These results are consistent with deep phylogenetic conservation of SR protein binding motifs. Assuming codons preferred near boundaries are "splice optimal" codons, in *Ectocarpus*, unlike *Drosophila*, splice optimal and translationally optimal codons are not mutually exclusive. The exclusivity of translationally optimal and splice optimal codon sets is thus not universal.

Key words: ESE, *Ectocarpus*, splicing, translational selection.

Introduction

Although for many years patterns of biased codon usage have been typically addressed in terms of translational optimality (and fit to the tRNA pool) (Duret 2002; Sharp et al. 2005), more recently the importance of exonic motifs involved in splicing has been seen to be relevant (Willie and Majewski 2004; Chamary and Hurst 2005; Parmley et al. 2006, 2007; Parmley and Hurst 2007; Warnecke et al. 2008). Chief among these motifs are exonic splicing enhancers (ESEs) (Blencowe 2000; Cartegni et al. 2002). At the RNA level, these motifs are responsible for the binding of SR proteins to the exonic parts of the unspliced RNA, thereby enhancing splicing at the neighboring exon–intron junction (Graveley 2000). In addition, they are responsible for retaining unspliced RNA in the nucleus (Taniguchi et al. 2007). Well-described ESEs in mammals—one of the few lineages where ESEs have been experimentally confirmed (Fairbrother et al. 2002, 2004; Fairbrother, Holste, et al. 2004; Ke et al. 2011)—are enriched toward the ends of exons (Fairbrother, Holste, et al. 2004), cause selective constraint at synonymous sites (Carlini and Genut 2006; Parmley et al. 2006), and have a highly skewed nucleotide usage,

typically being highly purine enriched (Tanaka et al. 1994; Fairbrother, Holste, et al. 2004; Parmley et al. 2007). Well-described ESEs occupy on average 30–40% of sequence near exon ends in mammals (Parmley et al. 2006). Note that as ESEs appear to be functional up to approximately 70 nt from an exon end (Fairbrother, Holste, et al. 2004), exons shorter than 140 bp can be considered to be all exon "end."

Owing to these three properties (high density, proximity to boundaries, and skewed nucleotide content), ESEs leave a marked footprint of codon usage near exon ends of mammalian genes, with codons more commensurate with involvement in ESEs (Parmley et al. 2007) being preferred near boundaries (Parmley and Hurst 2007). Similarly, when comparing synonymous codons, the one used more in ESEs is relatively preferred at exon ends over the synonym (Willie and Majewski 2004; Parmley and Hurst 2007). Thus, in mammals, although isochore composition is a strong driver of between-gene codon usage bias (Eyre-Walker and Hurst 2001), selection to preserve ESEs explains many of the intra–exon trends in codon bias. Amino acids also show skews in their usage as one approaches exon–intron junctions, with trends being well predicted by nucleotide content of ESEs and the codons that contribute to any given amino acid (Parmley et al. 2007). Indeed, comparing the usage of the 2-fold blocks of leucine and arginine with their respective 4-fold blocks supports the view that these trends are both the result of nucleotide-level effects and dominantly caused by splice-related constraints (Parmley et al. 2007). Just as knowing about ESEs makes sense of codon and amino acid trends, so too, conversely, *k*-mers that are enriched toward the ends of exons can be used to infer nucleotide preferences of splice-related motifs and to determine novel motifs (Lim et al. 2011) (N.B. codons are in frame 3-mers).

The trends seen in mammals have a series of further properties. For example, when usage trends at the 5′ and 3′ ends of exons are considered separately, it appears that the trends are largely symmetrical (Warnecke et al. 2008; Lim et al. 2011). That is, if a codon or amino acid is highly preferred at the 5′ end of exons, it is similarly highly preferred at the 3′ end. The logic of this symmetry is unclear, but it may accord with a model in which SR proteins aggregate on the ends of exons within the immature RNA and this aggregate defines, by the end of the cluster, a domain where the splice junction must reside. In such a model, there is no evident reason why different SR proteins should be under selection to bind 3′ and 5′ ends differently. However, such symmetry has to date only been observed in animals (Warnecke et al. 2008) and not in all of them. The 5′ ends of exons in *Caenorhabditis* worms, for example, are not simply different in composition to the 3′ ends; they show the opposite trends, that is, codons preferred at the 5′ ends are avoided at the 3′ ends and vice versa (antisymmetry). The 3′ end trends accord with the trends seen in all other taxa, with classical purine loading. The exceptional nature of worm's 5′ ends was hypothesized to reflect consequences of operonization in worm and the commensurate transplicing. The need to distinguish the 5′ ends of exons from the 5′ ends of genes, cut during transplicing, is suggested as the potential cause (Warnecke et al. 2008).

More generally, the trends in codon usage at the ends of exons in mammals correlate well with those seen in other animals, for example, *Drosophila* (Warnecke and Hurst 2007). This observation is important because *Drosophila*, unlike mammals, also has evident selection for use of "translationally optimal" codons, possibly to ensure mistranslation minimization (Akashi 1994; Drummond and Wilke 2008; Warnecke and Hurst 2010). In part, the cause of the strong correlation between end of exon usage in *Drosophila* and mammals reflects the fact that the "splicing optimal" set of codons and the "translationally optimal" set of codons are two almost mutually exclusive sets of codons, that is, translationally optimal codons tend to be those avoided near exon boundaries (Warnecke and Hurst 2007). At first sight, this mutual avoidance of the two sets seen in *Drosophila* makes some sense. If the two sets were the same, in highly expressed genes SR proteins would have difficulty binding exclusively to exonic ends, as all codons would be both translationally and splice optimal. Hence one might expect considerable splice disruption. Given such logic, it is worthwhile asking whether the same exclusivity rule applies in a very distantly related species.

Beyond *Drosophila*, whether the trends as observed in mammals are well conserved remains unclear, as the tendency to use SR proteins covaries with the intron density and size of introns (Warnecke et al. 2008). This trend possibly reflects an increased tendency toward exonization of splice site recognition as introns get ever larger, with small introns in a sea of large introns being the hardest to correctly splice using intronic information alone. At the other limit, a species such as *Saccharomyces cerevisiae* shows no preference trends (Warnecke et al. 2008), largely lacks SR proteins (Plass et al. 2008), and has very few and small introns. The nonanimal species previously analyzed (such as *Arabidopsis*) have very small introns and probably do not commonly use ESEs too, although SR proteins are possibly relatively ancient within eukaryotes but poorly described outside of the animal–fungal–plant crown group (Plass et al. 2008).

To examine whether the patterns seen in mammals might be relatively ancient requires analysis of distant genomes with abundant and relatively large introns. To this end, we selected for scrutiny the unusual genome of the brown alga *Ectocarpus siliculosus.* Brown algae share a common ancestor with the animal–fungal–plant crown group that predates the animal–fungal–plant common ancestor (Adl et al. 2005). The genome is well sequenced and annotated (Cock et al. 2010, 2012). It is unusual in being a nonvertebrate that is rich in introns (5.1 introns per kb of exon), and those introns tend to be large (mean intron size = 776 bp), meaning the genome is a strong candidate for one using ESEs and SR proteins to aid splicing,

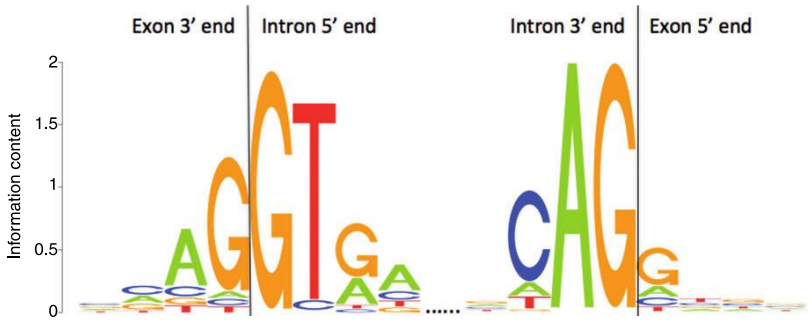


FIG. 1.—Splice site composition in *Ectocarpus*.

with a mean CDS size-to-gene size ratio of 0.27, comparable with mammals (Warnecke et al. 2008). As expected, annotation of the genome suggests it has SR proteins (Cock et al. 2010) (discussed later). The classical GT–AG rule applies in 95.3% of introns, the remainder being GC–AG introns (for sequenceLogo motifs, see fig. 1; for a longer span and evidence of a classical intronic 3′ polypyrimidine track, see supplementary fig. S1, Supplementary Material online). Importantly, much as with humans and other intron-rich genomes, but unlike some protists and intron-poor genomes (Irimia et al. 2007), there is not one hexameric motif that dominates intronic 5′ ends (GTGAGT at 12.5% is the most common). It thus appears an ideal candidate to ask whether the trends well resolved in humans are ancestral or animal specific. We also demonstrate that *Ectocarpus* has "translationally optimal" codons and thus ask whether these codons are never splice optimal codons.

Finally, taking advantage of what we discover to be some unusual features of the *Ectocarpus* genome, we reexamine the cryptic splice site avoidance model (Eskesen et al. 2004). This model posits that, with introns starting GT and exons ending in G, GGT should be avoided at the 3′ ends of exons (Eskesen et al. 2004) compared with the synonym GGC. *Ectocarpus* provides an unusually "clean" test of this prediction.

Materials and Methods

Establishing the Data Set for Analysis

The coding sequences (CDS) file and EMBL format exon information files for the brown alga *E. siliculosus* were downloaded from the database (http://bioinformatics.psb.ugent.be/genomes/view/*Ectocarpus*-siliculosus, last accessed September 16, 2013). The input CDS data were filtered to eliminate dubious sequences. We eliminated coding sequences that did not start with ATG, did not finish with a stop codon (TAA, TAG, and TGA), had internal stop codons, were not a multiple of three long, or contained one or more ambiguous nucleotides ("N"). In addition, those where the gene sequence length does not match the sum of the length of its exons as specified in the accompanying annotation files were eliminated. As we are interested in splice-related constraints, gene sequences that did not contain introns were also not examined. There are 16,579 coding sequences in the input file, of which 16,033 sequences qualified as suitable candidates.

Information of Expression Level of *Ectocarpus* Genes

The EST database of *Ectocarpus* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=%22Ectocarpus±siliculosus%22[porgn%3A__txid2880], last accessed September 16, 2013). Using BLAST, we identified the number of ESTs associated with each gene (identity $> 95\%$, e value < 0.01). The length-corrected EST hit rate (EST hits divided by the length of the gene) of each gene was regarded as the relative expression level of the gene.

HMMER Search for and Classification of SR Proteins

An SR protein reference data set, comprising 213 SR protein genes from different species, was established with the information from the website: http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node1.html (last accessed September 16, 2013). HMMER (Eddy 1998) was used to search for putative SR proteins in *Ectocarpus* genes (including those without introns) after multiple sequence alignment by MUSCLE.

To infer which, if any, of a set of cross species–conserved SR proteins our candidates might belong to, we performed a domain-based analysis, as previously described (Plass et al. 2008). In brief, we examined nine groups (families) of known SR proteins: SRp20 9G8, p54 SRp86, RY1, SC35-alias SRP1, SRm300, SRp30c-ASF, SRp40-55-75-alias SRP2, Topol-B, and Tra2. These were downloaded from http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node6.html (last accessed September 16, 2013). We aligned, using MUSCLE, the different groups of proteins

separately. We then used "hmmbuild" of HMMER to make an "hmm" profile for each multiple sequence alignment. All profiles were collected to form a profile database. Using "hmmscan," we searched all candidate *Ectocarpus* proteins against the profile database. Finally, we determined the SR protein family that best matched each *Ectocarpus* SR candidate. To this end, we considered those domains within a given *Ectocarpus* protein that are in the same order as in the reference SR protein (these being the "collinear" domains). We then summed the score of collinear domain hits for any given *Ectocarpus* protein for each reference SR protein. To choose which family a given *Ectocarpus* protein belongs to, we selected the one whose sum score of collinear domain hits was highest. Finally, we accepted this classification if the sum score of the collinear hits for a multi domain protein, or a single hit for a single domain protein, was equal to or greater than 100.

Determining Trends in Amino Acid and Codon Usage

According to the information in the EMBL annotation files, we extracted every exon sequence for every qualifying gene. The trend in usage of each codon and amino acid was investigated as a function of the distance from the exon–intron boundary up to a distance of 34 codons (to accord with an earlier analysis [Warnecke et al. 2008]). Importantly, the codon in direct proximity to the boundary was eliminated, but was used to analyze splice site profiles. The 5′ and 3′ ends were considered separately. The first and last exons were excluded, leaving 95,331 exons. For each codon and amino acid under consideration, we determined the slope on the line of proportional usage across all exons, as a function of distance from the boundary and the Spearman rank correlation (ρ). A negative slope on the line, or a negative ρ, indicates a codon or amino acid that is preferred near exon ends, whereas a positive slope implies a codon or amino acid preferred at exonic cores and avoided at the ends. In previous analyses, codons preferred near exon ends were well predicted by the composition of experimentally defined ESEs (Parmley and Hurst 2007).

Human Exonic Splice Enhancer Data Sets

The majority of systematic attempts to define human ESEs use computational approaches, confirmed with experimental support. Typically, these approaches start with a presumption about that distribution of ESEs and look for the sequences most enriched in these trends. We analyze three such data sets. Fairbrother et al. (2002, 2004) presumed that ESEs will be enriched in exons compared with introns and more abundant in exons with weak splice sites than in those with strong splice sites. This is the RESCUE-ESE data set. Zhang and Chasin presumed ESEs will be enriched in internal noncoding exons of protein coding genes compared with unspliced pseudo-exons and 5′ untranslated regions. This is the PESE data set. Goren et al. (2006) looked for motifs that were more conserved than expected at synonymous sites and enriched compared with background codon usage rates. This is the ESR data set. In the latter case, a minority of the motifs were exonic splice inhibitors, the precise proportion being uncertain not least because ESEs can also function as exonic splice inhibitors depending on their position and context within the exon (Ke et al. 2011). The fourth data set we consider, Ke-ESE, derives from a purely experimental approach adopted by Ke et al. (2011). They considered the effects of all possible 4,096 6-mers at five locations in two model exons. Taking into account overlap sequences, this permitted the identification of numerous ESE hexamers.

We downloaded the ESR and Ke-ESE hexamers directly from the original papers. For the Ke-ESE set, we selected, as the authors did, the 400 hexamers with the highest scores. The RESCUE-ESE data set was downloaded from http://genes.mit.edu/burgelab/rescue-ese/ESE.txt (last accessed September 16, 2013), and the PESE original octamers were downloaded from http://www.columbia.edu/cu/biology/faculty/chasin/xz3/pese262.txt (last accessed September 16, 2013). PESE hexamers were extracted from octamers with a minimum of seven occurrences.

Assembling a Set of *Ectocarpus* Putative ESEs

The attempts to infer human ESEs have, as noted earlier, typically specified two criteria whereby ESEs are expected to be enriched (i.e., a ***R***elative ***E***nhancer and ***S***ilencer ***C***lassification by ***U***nanimous ***E***nrichment = RESCUE method). Here, we perform a similar RESCUE approach to define *Ectocarpus* ESEs. We consider that ESEs should be 1) enriched at exonic ends compared with introns and 2) that the usage of the ESE should increase from exon core to exon flank.

To determine the latter, for all 4,096 possible hexamers we considered their relative usage in exons, in all frames, as one moves away from exon ends. We considered only those exons longer than 160 bp to ensure that enrichment at exonic ends is truly such enrichment, rather than enrichment in short exons. The 5′ and 3′ ends were considered separately.

To consider those hexamers enriched at exon ends compared with intronic sequence, we considered exons longer than 100 bp and introns longer than 100 bp. We then considered the terminal 50 bp at each end of the exons and 50 bp at the end of the introns. For statistical analysis, it is important that there are the same number of introns as exons, so we randomly sampled from the larger data set to equalize the size of the two.

For each hexamer, we then considered its mean usage at exon ends and its mean usage at intron ends. We then calculated the difference in usage between the exon end and intron end. The 5′ exonic ends were compared with the 3′ intronic ends and vice versa. For each hexamer we can then define

$$\delta_{\text{Observed}} = \text{Exonic density} - \text{Intronic density}.$$

To consider the significance of this, we then pooled the relevant data from exons and introns, randomized them, and then considered the first half of the data as being pseudo-exon and the second half pseudo-intron. Repeating this 100 times for each hexamer we define

δ_{Pseudo} = Pseudo exonic density − Pseudo intronic density.

A reasonable metric of the extent to which a given hexamer is enriched at exon ends compared with intronic ends is then:

$$Z = \frac{\delta_{\text{Observed}} - \overline{\delta_{\text{pseudo}}}}{\sigma_{\delta\text{Pseudo}}}$$

where $\overline{\delta_{\text{pseudo}}}$ is the mean of the hexamer usage in the 100 pseudo sets and $\sigma_{\delta_{\text{Pseudo}}}$ is the standard deviation in the usage across the pseudo sets. P was approximated by extrapolation from Z under an assumption of normality.

To generate a set of ESEs, we then considered those hexamers enriched in exon end compared with intron ($Z > 0$) and preferred near exon ends compared with core ($\rho < 0$, slope < 0), and then combined P values from the two approaches using Fisher's method. We then considered those hexamers with a combined $P < 0.05/4{,}096$ as putative ESEs.

CAI Calculation, Identification of "Optimal" Codons, and the Relationship with Gene Expression

A data set containing 43 ribosomal proteins was established and used as a reference "highly expressed" gene class. The codon usage in this set was analyzed using CodonW. We used this reference data set and the reference codon usage table from Codon Usage Database (www.kazusa.or.jp/codon/countcodon.html, last accessed September 16, 2013) to determine codon adaptation index (CAI) scores for all genes. To this end, we downloaded the local version of CAIcal, a CAI calculation program, from http://genomes.urv.es/CAIcal/ (last accessed September 16, 2013) and calculated the CAI values of all genes, except the ribosomal genes, which had been used for establishing the reference data set. The validity of the CAI index was examined by considering the relationship between expression level and CAI, again excluding the ribosomal genes. For any given synonymous block, the codon with the highest codon adaptation index was considered to be the "optimal" codon. tRNA copy numbers were obtained from http://plantrna.ibmp.cnrs.fr/ (last accessed September 16, 2013).

Alternative Splicing Event Calculation

Alternative splicing events in human and *Ectocarpus* genes were identified from 8,315,122 and 67,082 EST sequences, respectively, downloaded from the dbEST database (Boguski et al. 1993), using methods previously outlined (Chen et al. 2011). In brief, individual ESTs were matched to individual genes by aligning them to the genome sequence using GMAP (Wu and Watanabe 2005). Exon templates were then inferred from EST alignment coordinates. Alternative splicing events were identified by comparing alignment coordinates for each EST against the exon template.

Comparable alternative splicing event counts correcting for EST coverage were obtained using a transcript normalization protocol as described previously (Kim et al. 2007), where alternative splicing events per gene are calculated as the average number of alternative splicing events identified in 100 random samples of 10 ESTs.

Results

Ectocarpus Has Multiple SR Proteins

Before asking whether binding of SR proteins to ESEs leaves a footprint of biased codon and amino acid usage in proximity to intron–exon boundaries, we first established the profile of SR proteins within the genome. To search for candidates, we did a HMMER search, training the HMMER on an established collection of SR proteins. In total, we identified 54 putative SR proteins, including three previously annotated as SR proteins (Esi0638_0002, Esi_0327_0029, and Esi0164_0021) (supplementary result S1, Supplementary Material online). In the original build of the genome, a further putative SR protein was identified (Esi0089_0034; annotated as "splicing factor, arginine/serine-rich 2, RNAp interacting protein, putative"). This was not identified by HMMER. Although many of the extra hits are unlikely to be SR protein (e.g., eukaryotic translation initiation factor 3′, subunit a), several more have suggestive RNA binding functions. Most of the extra hits are not annotated.

To clarify just which SR proteins the HMMER search might have revealed, we performed an additional domain-based analysis, comparing our proteins with nine SR proteins that are relatively well conserved through plants, animals, and fungi (Plass et al. 2008). We found robust evidence for 18 of the *Ectocarpus* genes being members of 1 of 5 SR protein families (fig. 2; supplementary table S1, Supplementary Material online). This includes the three previously annotated SR proteins (supplementary table S1, Supplementary Material online). We conclude that *Ectocarpus* has a good number of SR proteins but probably not the full set described in humans (Long and Caceres 2009).

A High Proportion of Amino Acids and Codons Show Preference/Avoidance Trends

To determine which, and how many, codons and amino acids show significantly skewed usage in proximity to exon–intron junctions, we considered the relative usage of all codons and amino acids as a function of distance from an exon–intron junction, ignoring the codon in immediate proximity to the junction. We examined the 3′ and 5′ ends separately. Any codon or amino acid preferred near a boundary will have a negative slope and a negative Spearman rank correlation (ρ)

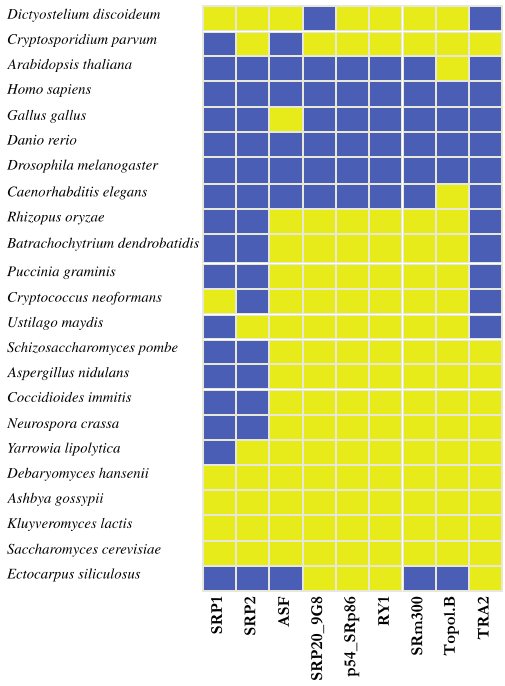


FIG. 2.—Presence or absence of SR and SR-related proteins in *Ectocarpus* compared with other taxa. Presence of at least one member of a gene family in a given species is indicated in dark blue or absence in light yellow. Data for all species bar *Ectocarpus* from Plass et al. (2008).

between its relative usage and the distance from the boundary. A positive slope/ρ score indicated avoidance near a boundary relative to usage more core to exons. The slope/ρ values, we considered to be measures of the preference/avoidance trends.

Most codons and amino acids showed significant preference/avoidance directions near exon–intron boundaries (supplementary tables S2 and S3 and figs. S2 and S3, Supplementary Material online). Before Bonferonni correction, 86% of codons and 96% of amino acids showed significant trends ($P < 0.05$) at the 5′ exonic ends, and 88% of codons and 91% of amino acids showed significant trends ($P < 0.05$) at the 3′ exonic ends. After correction, these numbers dropped to 68% of codons and 83% of amino acids showing significant trends at the 5′ exonic ends, and 69% of codons and 83% of amino acids showing significant trends at the 3′ exonic ends. Overall, considering all codons with at least one synonym, 66% of comparisons showed significant trends after Bonferonni multitest correction, and 83% of amino acids analyses showed significant trends.

These figures compare strikingly with what has been seen before. A priori we expect that species with relatively small

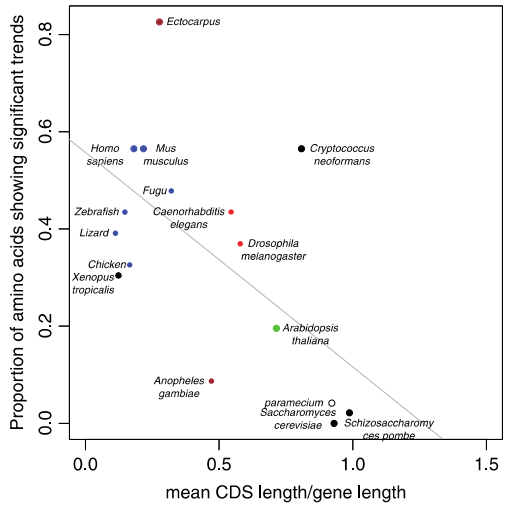


FIG. 3.—The proportion of amino acids showing significant preference/avoidance trends after Bonferroni correction as a function of the average ratio of mature CDS to gene length across multiple species.

exons sitting in an intron-rich sea will be those that will be under selection for adding exonic splice information to bolster the intronic signals. This should in turn be reflected in more codons and more amino acids showing skewed usage near boundaries. This supposition is generally supported by the finding that species in which the ratio of the mature CDS size to gene size is small are those in which a greater proportion of amino acids or codons show significant skews toward exon ends (fig. 3). When we consider our new data in this light, while *Ectocarpus* certainly has a low CDS-to-gene ratio, the proportion of amino acids showing a skew remains unusually high (fig. 3). We note that this cannot be an artifact of sample sizes (the higher the sample size, the more likely significant skews will be seen even if trends are weak), as humans and mice have fewer significant trends but more exons analyzed.

Preference/Avoidance Trends at 3′ and 5′ Exon Ends Are Similar

Is the pattern of symmetry seen in most animals, but not so far reported outside of animals, also seen in *Ectocarpus*? To address this, we considered for each amino acid and each codon the trend in its usage approaching the 5′ and 3′ ends of exons. The slope on this line and the Spearman ρ values were considered. We then considered the correlation between the figures when comparing the 5′ and 3′ ends. We found that overall exons tend to be symmetric, with a strong correlation

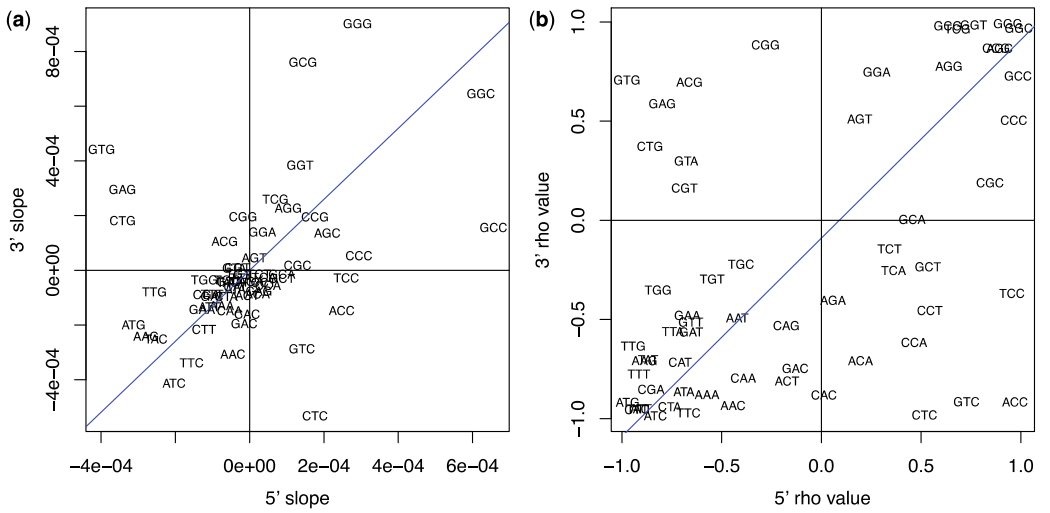


Fig. 4.—Examination of symmetry of preference/avoidance trends for codons. For both 5′ and 3′ exon ends, we considered (*a*) the slope on the line of relative usage versus distance from the boundary ($\rho = 0.40$, $P = 0.0014$) and (*b*) the Spearman rank correlation for the same comparison ($\rho = 0.47$, $P = 0.00014$). A negative slope or a negative ρ indicates a codon that is preferred near an exon boundary. For each codon, we can then compare these trends at the 5′ and 3′ ends. We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

both between the slopes and the ρ values for codons (fig. 4: from binomial test with success as preservation of the sign of the slope, $P = 0.004$; from Spearman rank correlation on slopes, $P = 0.001$) and amino acids (fig. 5: from binomial test with success as preservation of the sign of the slope, $P = 0.0026$, from Spearman rank correlation on slopes, $P = 0.002$).

However, closer scrutiny indicates a further nuance, namely, that C and G ending codon usage can be antisymmetrical (fig. 4). CTC, GTC, and ACC are all highly disfavored at exonic 5′ ends but highly preferred at 3′ ends, whereas GTG, GAG, ACG, CGG, and CTG all show the opposite pattern. Generally, at 4-fold degenerate sites, C and G show somewhat antisymmetrical patterns, with C avoided at 5′ ends while G, although highly abundant, avoided at 3′ ends, meaning as one approaches the boundary, its usage declines (fig. 6*a* and *b*). However, these patterns are not simple enough to be explained solely in terms of C or G content, as many C ending codons, including CCC, are symmetrical. Nonetheless, we conclude that the symmetry rules are not universally respected.

Usage Trends Near Exon Junctions in *Ectocarpus* Resemble Those Seen in Humans

To consider whether the trends seen in humans are also seen in a very distant species with large intronic content such as *Ectocarpus*, we considered how the ρ values for each amino acid, 5′ and 3′ resembled those seen in humans. We can then ask two questions. First, do the trends overall correlate between the two species? Second, what proportion of comparisons has a conserved direction of preference/avoidance between the two species?

To this end, we considered the correlations between the trends (ρ values) seen in the two species, considering 5′ and 3′ ends separately. Remarkably, despite over 1 billion years of divergence, we found that at both 5′ and 3′ exonic ends the trends correlate well (5′ end: $\rho = 0.68$, $P = 0.0005$; 3′ end: $\rho = 0.53$, $P = 0.01$; fig. 7 and table 1). At the 3′ ends, only 4 of the 23 codon blocks (we treated 2-fold and 4-fold blocks differently here) did not have a conserved preference trend (binomial test, $P = 0.002$). At the 5′ ends, although the trends are correlated, seven codon blocks had different trends, rendering the result nonsignificant (binomial test, $P = 0.09$). However, when we restricted analysis to those amino acids showing significant trends (before Bonferroni correction), we observed that at both the 5′ and 3′ ends, the proportion conserved was significant (table 1). Moreover, we noted that of the four amino acids that were antisymmetrical in *Ectocarpus* (E, R, V, and T), three (E, V, and T) were also antisymmetrical in humans (supplementary fig. S4, Supplementary Material online).

We can perform a similar analysis using codons (excluding ATG and TGG owing to lack of synonyms). Again we found strong concordance between trends seen in humans and those seen in *Ectocarpus* (5′ end: $\rho = 0.50$, $P = 5.17 \times 10^{-5}$; 3′ end $\rho = 0.58$, $P = 1.7 \times 10^{-6}$; fig. 8 and table 1). The conservation patterns mirrored what we saw at the amino acid

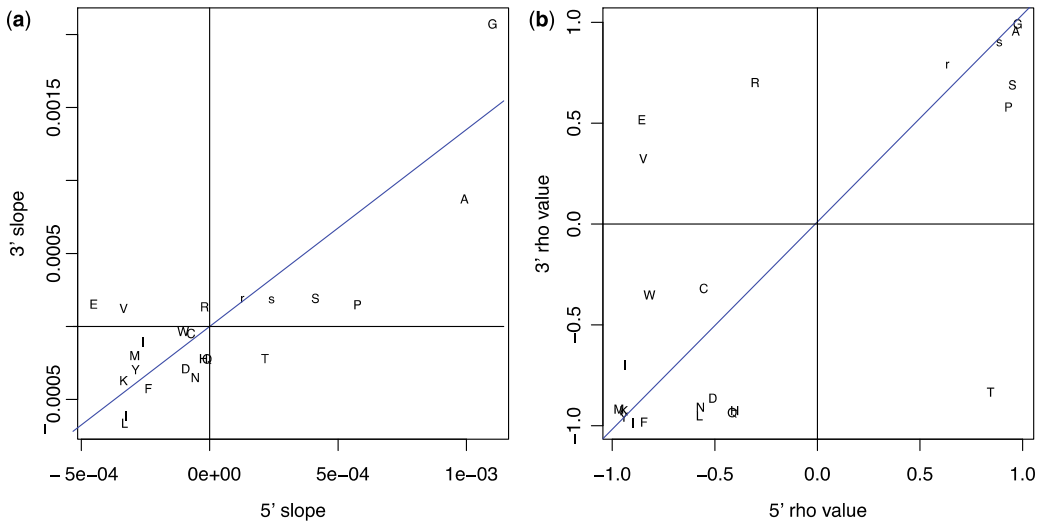


FIG. 5.—Examination of symmetry of preference/avoidance trends for amino acids. For both 5′ and 3′, we considered both the slope on the line of relative usage versus distance from the boundary and the Spearman rank correlation for the same comparison. For each amino acid, we can then compare these trends at the 3′ and 5′ ends, considering either (*a*) slope ($\rho = 0.60$, $P = 0.003$) or (*b*) ρ ($\rho = 0.68$, $P = 0.0005$). We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

level. At the 3′ ends, we saw a strong correlation, and only 12 showed reverse trends (binomial test, $P = 5.13 \times 10^{-6}$). At the 5′ ends, the effects were more modest. A significant correlation was observed, but of 59 codons, 23 showed reverse trends at the 5′ ends ($P = 0.12$). As above, restricting analysis to only those codons showing significant trends, at both 5′ and 3′ ends, more than expected show conservation of direction (table 1). Nucleotide usage at 4-fold degenerate sites was also comparable between *Ectocarpus* (fig. 5*a* and *b*) and humans (fig. 5c and *d*), although in *Ectocarpus* the C and T preferences at the 3′ end were more similar than seen in humans. Overall, these results suggest a deep phylogenetic conservation of splice-associated trends in amino acid composition as one approaches exonic ends, most especially at the 3′ ends.

Ectocarpus Has Low Rates of Alternative Splicing

The similarity that we see between humans and *Ectocarpus* in terms of which codons and amino acids are preferred and avoided near boundaries suggests that the selection on the nucleotide usage in DNA or RNAs at exon ends is for similar reasons. Why then does *Ectocarpus* have so many more amino acids and codons showing significant trends (fig. 2)? The metric we use is by no means perfect, as it is sensitive to sample sizes (number of exons examined). However, high numbers of trends seen for *Ectocarpus* compared with mouse/human cannot be an artifact of sample sizes, as the sample sizes in vertebrates (in terms of number of exon ends) are larger than those in *Ectocarpus*.

One possibility that explains the large number of skews in *Ectocarpus* is that alternative splicing might be relatively rare in *Ectocarpus*. The consequence of this would be that most exons are consistently under strong selection to be spliced correctly. By contrast, if in humans many exons are splicing errors (Zhang et al. 2009), then we would not expect strong selection to preserve ESEs in all exons. The uniformity of splice sites in *Ectocarpus* (fig. 1) would be consistent with the hypothesis that most exons are under selection to be properly spliced.

Preliminary data suggest that alternative splicing is indeed rare in *Ectocarpus*. A detailed examination of splice forms has been performed on one gene family, the cytosolic glutathione transferases. While 11 genes were identified, only one had an alternative transcript (Franco et al. 2008). Although this is much lower than the rate seen in humans, in whom nearly all intron-bearing genes have at least two isoforms (Pan et al. 2008), this difference might reflect, at least in part, differences in the depth of study (Brett et al. 2002).

To compare alternative splicing levels in both humans and *Ectocarpus* allowing for depth of EST sequencing, we performed two tests. First, we measured alternative splicing levels in genes from both species after transcript number normalization. For this, alternative splicing per gene was measured as the average number of alternative splicing events detected in 1,000 random samples of 10 ESTs. We obtained

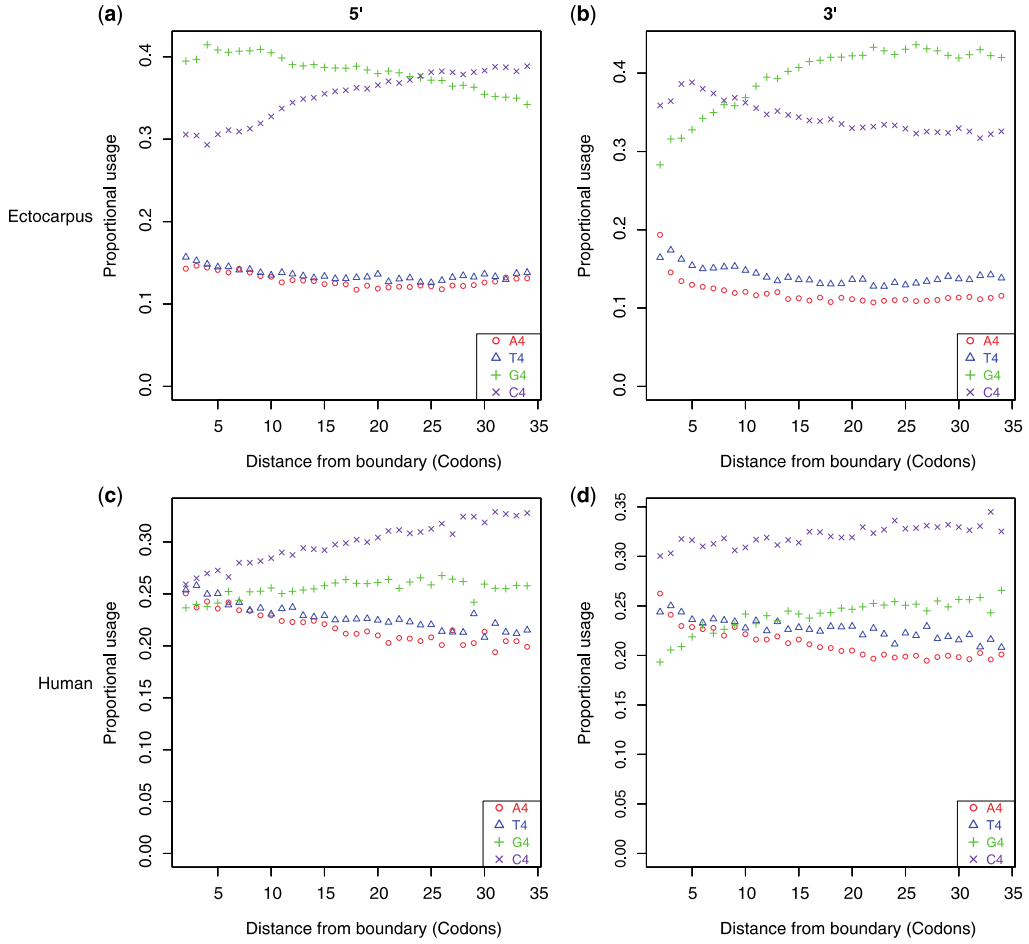


FIG. 6.—Nucleotide usage at 5′ and 3′ exon ends at 4-fold degenerate sites in *Ectocarpus* and humans. The data here use only exons longer than 64 codons so that all exons contribute equally at all distances. The plots are (*a*) *Ectocarpus* 5′ end, (*b*) *Ectocarpus* 3′ end, (*c*) human 5′ end, and (*d*) human 3′ end.

this comparable index of alternative splicing for 8,772 human genes and 69 *Ectocarpus* genes. We found that while *Ectocarpus* genes had an average of 0.41 events per gene (median of 0), human genes had an average of 5.35 events per gene (median of 4.55). This difference is highly significant (*t*-test, $P = 2.15 \times 10^{-68}$). Second, we compared the average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. When genes were divided according to their average number of aligned ESTs, the average number of alternative splicing events per gene was considerably higher for humans compared with *Ectocarpus* at all nine EST per gene counts ($P = 0.004$ from binomial test: $N = 4{,}861$ human; 326 *Ectocarpus*, fig. 9). We conclude that in *Ectocarpus* alternative splicing is rare compared with that seen in humans. Although this is consistent with the possibility that alternative transcription rates might impact on the net skew in nucleotide usage, this hypothesis requires considerable further cross-taxon analysis.

Ectocarpus Putative Exonic Splice Enhancers Resemble Those Seen in Humans

Above we compared human and *Ectocarpus* exonic ends as regards trends in codon usage. The trends seen in

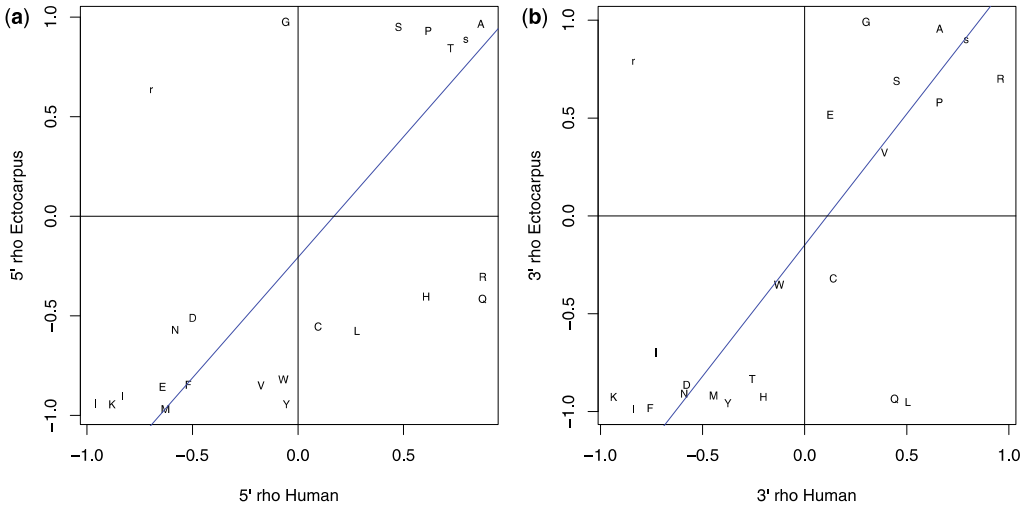


Fig. 7.—Comparison of preference/avoidance trends at the amino acid level between humans and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by ρ (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at (*a*) 5′ ($\rho = 0.68$, $P = 0.0005$) and (*b*) 3′ ($\rho = 0.53$, $P = 0.01$) ends of exons are shown. The blue line is the SMA regression line.

Table 1
Conservation of Trends between Humans and *Ectocarpus*

	Binomial Test		Spearman's Rank Correlation	
	Changed Direction	*P*	ρ	*P*
All observations				
5′ AA	7 from 23	0.093	0.6779	0.0005
3′ AA	4 from 23	0.0026	0.5316	0.0100
5′ codon	23 from 59	0.1175	0.5017	5.17E−05
3′ codon	12 from 59	5.13E−06	0.5773	1.70E−06
Significant observations				
5′ AA	3 from 16	0.021	0.7559	0.0011
3′ AA	3 from 17	0.013	0.5000	0.0430
5′ codon	11 from 43	0.0019	0.5694	6.75E−05
3′ codon	5 from 38	4.26E−06	0.5938	8.51E−05

Note.—The trends in usage of all codons and amino acids at 3′ and 5′ ends of exons were compared between humans and *Ectocarpus*. Two reporting statistics were considered. First, we ask using a binomial test whether the proportion of observations changing direction of trend is different from expected under a null where trends are free to evolve. Second, we consider a Spearman rank correlation test. As the analysis can be biased by considering trends that are very marginal and nonsignificant, we perform a second analysis where only significant trends ($P < 0.05$ before Bonferroni correction in both species) are used.

mammals reflect the nucleotide content of ESEs (Parmley and Hurst 2007). This is to be largely expected, as ESEs are hexamers that tend to be enriched at exon ends in any frame and codons are 3-mers in frame and hence are likely to be nonindependent of ESE-imposed trends. ESEs are, however, not described in *Ectocarpus* so we cannot perform the same analysis. We can, however, attempt to determine which hexamers might function as ESEs and compare this set of candidates with those identified in humans.

To this end, we asked of *Ectocarpus* 1) which hexamers are enriched at exonic ends compared with intronic sequence and 2) which hexamers are used at exon ends more than in exon centers (i.e., have a negative slope of proportional usage against distance from exonic end). We then considered as candidate ESEs the hexamers most enriched on both axes. Note that this method is far from perfect in so much as we also identified causes of skew in codon usage probably not related to ESEs but rather to avoidance of cryptic splice sites (discussed later).

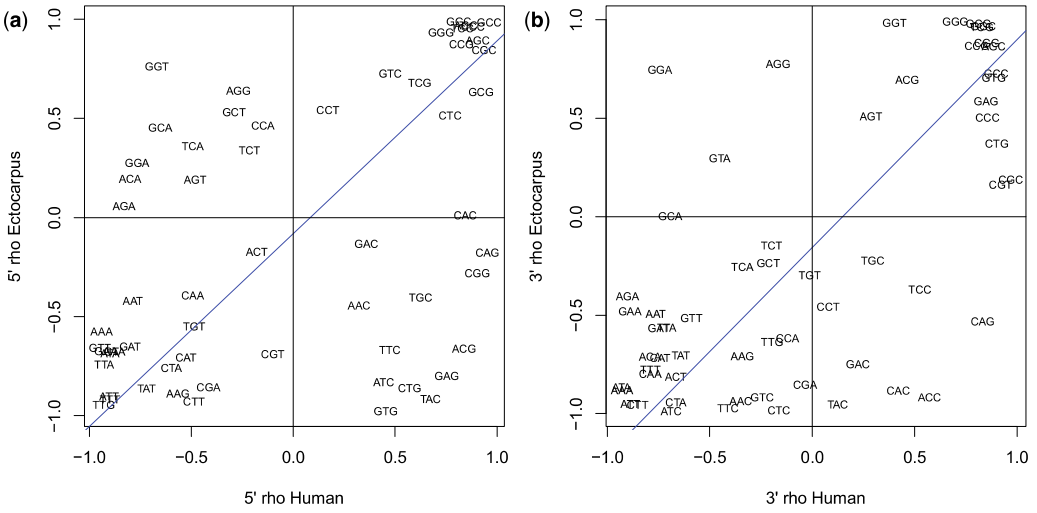


FIG. 8.—Comparison of preference/avoidance trends at the codon level between humans and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by ρ (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at (*a*) 5′ ($\rho = 0.50$, $P = 5.17 \times 10^{-5}$) and (*b*) 3′ ($\rho = 0.58$, $P = 1.7 \times 10^{-6}$) ends of exons are shown. The blue line is the SMA regression line.

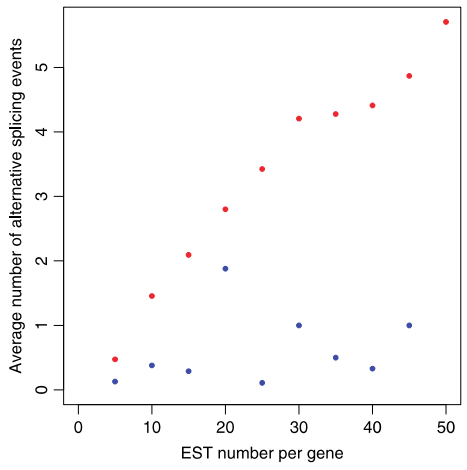


FIG. 9.—The average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. Data for humans in red; data for *Ectocarpus* in blue.

We identified 904 3′ ESEs and 919 5′ ESEs (supplementary table S4, Supplementary Material online). The 5′ and 3′ hexamers are different from each other. We observed 189 in common but by chance, we would expect 203. Moreover, while the 5′ hexamers are, like classically described SR protein-binding ESEs, highly purine enriched (A + G = 64.4%), the 3′ set are, if anything, pyrimidine enriched (A + G = 42%), which is consistent with the C enrichment at 4-fold sites at the 3′ exonic ends (fig. 6). The set of 189 ESEs that are common to 5′ and 3′ ESEs are purine rich (A + G = 59.3%).

There are four high-throughput data sets attempting to identify ESEs in humans (see Materials and Methods). Unfortunately, these four have remarkably few hexamers in common—just 10 of more than 900 putative hexamers are found in all four data sets. We consider those hexamers found in at least 3 of the 4 data sets as being a robust set of human ESEs ($N = 54$). For both our 3′ and 5′ set of hexamers, we found considerably more overlap than expected under a null model in which the human set of ESEs and the *Ectocarpus* set were assumed to be independent. For the 5′ ESEs, we expected 12 hexamers in common but observed 39, more than 8 standard deviations than expected by chance ($P \ll 0.0001$). For the 3′ end ESEs, the effect was more modest but still highly significant: we observed 26 in common between the two sets, where less than 12 are expected by chance, nearly 5 standard deviations than expected by chance ($P < 0.0001$). Of the 189 hexamers that are in common at 5′ and 3′ ends, 18 are also in the set of 54 human ESEs while fewer than 3 are expected under a random null. This deviation is almost 10 standard deviations from expectations ($P \ll 0.0001$). All of these degrees of concordance between *Ectocarpus* and humans are considerably

greater in magnitude than the concordance witnessed between some of the initial four human data sets. We conclude that despite the unusual base composition of 3′ ESEs in *Ectocarpus*, there is a significant resemblance between human ESEs and *Ectocarpus* ESEs. The trends, especially those seen at the 5′ ends, are consistent with a deep and strong phylogenetic conservation of SR protein-binding preferences.

Translationally Optimal and Splice Optimal Codons Are Not Mutually Exclusive in *Ectocarpus*

In *Drosophila*, the set of codons enriched near exon ends accords with those commensurate with ESEs, and correlate well with the trends seen in mammals (Warnecke and Hurst 2007). These splice optimal codons are very different from the translationally optimal codons, with just one codon being in both sets (Warnecke and Hurst 2007). Is this mutual exclusivity also seen in *Ectocarpus*? To address this, we first must ask whether *Ectocarpus* is like *Drosophila* in having a translationally optimal class of codons. To this end, we first examined codon usage in the ribosomal proteins, these being the most highly expressed genes. Given the difference in the codon usage in the ribosomal genes and the codon usage in the genome as a whole, we could then ascribe each gene a CAI score. We then ask whether, excluding the ribosomal protein training set, the more highly expressed genes show higher CAI. There is a weak but significant correlation between CAI and expression level (Pearson correlation, $r = 0.084$; $P = 2.6 \times 10^{-13}$, supplementary fig. S5, Supplementary Material online).

In addition, we compared the optimal codons, as defined by over usage in ribosomal proteins, for each synonymous set with the tRNA copy numbers (assuming these to be a rough guide to tRNA levels) and asked if the optimal codon within each block was also the one with the most abundant tRNA. In 12 of 18 synonymous blocks, this was the case (supplementary table S5, Supplementary Material online). By randomizations, involving extracting at random two codons from each synonymous block, we asked how often we expected by chance to see 12 of 18 matching, given the structure of the genetic code. In 100,000 simulations we observed 12 or more matches in less than 1,000 incidences. We conclude that the optimal codons tend to be those matching the more abundant tRNAs ($P < 0.001$). *Ectocarpus* is, in this regard, more like flies than mammals, and is under translational selection.

Given the above result, we can now address whether the translationally optimal codons might be different from the splice optimal codons. To define splice optimal codons, we consider all those preferred near exon boundaries (at both 5′ and 3′ ends) that are significantly skewed after Bonferonni correction ($P < 0.05/118$ at both ends and $\rho < 0$) (supplementary table S5, Supplementary Material online). This defines 16 codons, although some of these are from the same codon block. Indeed, of 18 amino acids with more than one synonym, 10 amino acids have no splice preferred codons. In the remaining cases, three amino acids have all their codons as splice optimal (F, I, K, and Y). In three (H, L, and R) of the remaining four informative cases the translationally optimal codon, defined by reference to usage in ribosomal proteins, is not a splice optimal codon, but in K it is. To examine the significance of this, we considered a simulation in which we define for each of the four codon blocks the number of splice optimal codons and randomly sampled that number out of the number of codons in the block. We then ask how often the pseudo-splice set of codons and the pseudo translationally optimal codon matches. We then considered how often we see 1 or fewer matches. We found that we expect this to happen about 41.6% of the time, thus there is no evidence that splice optimal and translational optimal codons are under selection to differ.

We can be less stringent and define a splice optimal codon as any codon showing preference toward any exon end (not both 5′ and 3′ ends) after Bonferonni correction ($P < 0.05/118$). This gives 34 splice optimal codons (supplementary table S5, Supplementary Material online). There are only four potentially informative synonymous codon blocks in which some but not all of the codons are splice optimal. As regards translational optimality defined in terms of usage in ribosomal proteins, N and T have splice optimal codons that are translational optimal ones, whereas Q and R have the opposite. Again we see no significant evidence that splice optimal and translational optimality are divergent (from simulation: $P = 0.65$).

Additionally, for each codon block we can ask which codon is the most splice preferred. This we define as the codon with the most significantly negative slope using both 5′ and 3′ analyses. If no codon has a significantly negative slope then we consider the one with the most negative slope to be the splice preferred codon. We find that in 9 of 18 incidences the splice preferred codon is also the translationally optimal codon. We reject the hypothesis that splice optimal codons tend not to be translationally optimal codons (by simulation: $P = 0.91$).

Evidence of Cryptic Splice Site Avoidance

The nucleotide composition at 3′ exonic ends, allows us to provide an unusually "clean" test of the cryptic splice site avoidance model (Eskesen et al. 2004). Given that introns start GT and end AG, to avoid cryptic splice sites, it is argued (Eskesen et al. 2004) that AG residues should be avoided near 5′ ends of exons and GT should be avoided at the 3′ ends. One difficulty with any such analysis in mammals, however, is that, nucleotide usage in ESEs trends to go in the same direction as predictions from the cryptic splice avoidance model (Chamary and Hurst 2005). *Ectocarpus* provides an opportunity to test the cryptic splice model as at the 3′ ends both T and C are weakly preferred and show very similar relative trends (fig. 6*b*). As exons tend to end G in the majority of

incidences (fig. 1), the cryptic splice avoidance model thus predicts that at 3′ exon ends GGT should be avoided compared with GGC (a cryptic splice could occur between the two G residues in GGT), but [A|C|T]GT need not be avoided compared with [A|C|T]GC. Precise expectations for [A|C|T]GC and [A|C|T]GT are not however clear, their relative usage potentially reflecting background nucleotide trends. Given this, we asked solely whether GGT/GGC behaves differently from [A|C|T]GC and [A|C|T]GT, with the latter three consistent in their behavior. We observed just this, with profound avoidance of GGT compared with GGC, but preference for [A|C|T]GT, compared with [A|C|T]GC, near boundaries at the 3′ ends of exons (fig. 10). Note that both GGN and CGN are 4-fold degenerate codons so this comparison is especially well controlled. At the 5′ ends of exons the pattern is reversed with GGT being preferred over GGC, which is to be expected given the overall nucleotide composition, C being strongly avoided 5′ and T being weakly preferred. In sum, the preference of GGC over GGT at exonic 3′ end is consistent with the cryptic splice site model.

At the exonic 5′ end as introns end AG and exons commonly start with a G (fig. 1), the cryptic splice model predicts that AGG should be avoided compared with AGA. This is observed (supplementary fig. S6, Supplementary Material online). This test is not a strong one however as, while exons regularly end G, the preference to start with a G is weaker.

Although the cryptic splice model makes good sense of the preference for GGC over GGT at 3′ ends, most other trends cannot be explained in terms of splice avoidance. For example, preference for [A|C]GT over [A|C]GC most probably reflects processes acting more generally. We presume, as typically done (Lim et al. 2011), that most of the trends observed reflect the nucleotide content of splice motifs, such as ESEs.

Discussion

The analysis of the *Ectocarpus* genome has provided the first insight into the splice-related forces operating in a very distant relative of vertebrates in a species with intronic content comparable with that of vertebrates. The extent to which the trends observed in vertebrates accord with those seen in *Ectocarpus* are striking given the vast evolutionary distance between the two groups. It seems, therefore, parsimonious to presume that this reflects splice-related constraints, most probably the conservation of the binding motifs of SR proteins, not least because trends seen in humans accord well with those expected given the nucleotide composition of ESEs (Parmley and Hurst 2007; Parmley et al. 2007). Moreover, that *k*-mer enrichment in the vicinity of exon boundaries is a successful method to identify new splicing motifs, supports the supposition that the codon trends that we observe reflect splice-related motifs (Lim et al. 2011). The correspondence between our predicted set of *Ectocarpus* ESE hexamers and human ESEs is notable. It is to be expected that, just as vertebrate ESEs can function in fungi (Webb 2005), so they might function in brown algae too. Nonetheless, given our results regarding the cryptic splice site avoidance model, we see no reason to suppose that ESE enrichment is the sole cause of all the trends that we observe.

Why *Ectocarpus* has so many codons and amino acids showing strong preference avoidance trends (and also so many putative ESEs) is unclear. The possibility that alternative splicing is rare in *Ectocarpus*, hence resulting in selection on most exons most of the time for correct splicing, is consistent the data but requires further scrutiny. As regards the trends seen at the amino acid level, an alternative possibility to splice related selection is that we are detecting preference for one amino acid above another, owing to the hypothesized tendency of protein modules to reside in individual exons, as conjectured by the introns-early hypothesis (Gilbert et al. 1986). Aside from the fact that the one-module one-exon hypothesis is probably untenable (Stoltzfus et al. 1994; Logsdon 1998), this possibility is rejected in humans, not least because of the 6-fold degenerate amino acids, two (L and R) show opposite trends within the 2-fold and 4-fold degenerate blocks, these trends being well predicted by involvement in ESEs (Parmley et al. 2007). Similarly, for the 2-fold block of arginine (R) in *Ectoparpus*, at the 5′ exon ends both codons are avoided, whereas the three of the four codons of the 4-fold block are preferred. Within the 4-fold blocks of both valine and threonine there are both codons that are significantly avoided and significantly preferred. Similarly, within the 4-fold degenerate block of arginine and the 2-fold degenerate glutamic acid at 3′ exon ends one of the two is significantly avoided and one is significantly preferred. These differing trends within synonymous codon blocks support the hypothesis that, at least in part, the trends that we observed are owing to nucleotide level, not protein level, effects. Nonetheless, most pairs of codons in 2-fold degenerate blocks have preference trends in the same direction. This could reflect either some relationship to protein structure or similar splice-related selection (e.g., ESE involvement) owing to the first two bases in the 2-fold degenerate codons being identical in the synonyms.

Unlike *Drosophila*, *Ectocarpus*, while having evidence of being under translational selection, shows no evidence of selection to make splice optimal and translationally optimal codons distinct. Why might the two genomes differ? One possibility is that selection for translational optimality is that stronger in *Drosophila*. Indeed in *Ectocarpus* the correlation between CAI and expression level is rather weak. Were this weakness real (as opposed to an artifact of limited and noisy expression data) then selection to force divergence between translationally optimal and splice optimal codons may also be weak. Another possibility is that the observation

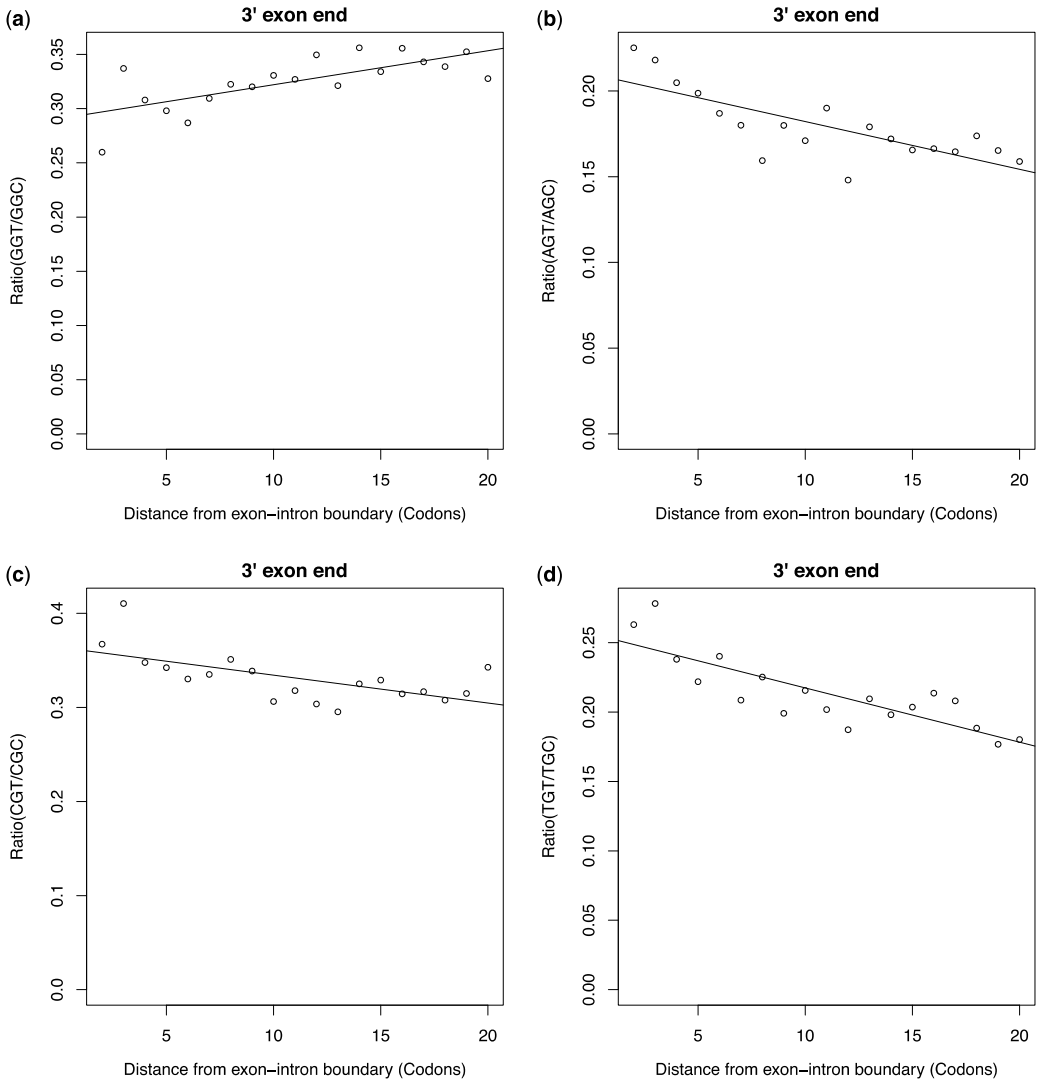


Fig. 10.—Relative usage of NGT against NGC at synonymous sites at exonic 3′ ends (*a*) N = G, (*b*) N = A, (*c*) N = C, (*d*) N = T.

in *Drosophila*, while seemingly having an attractive explanation, is an accidental consequence of selection on translational optimality and splice optimality happening to go in opposite directions. Until it is better understood why certain codons end up being translationally optimal this issue will be hard to resolve. Nonetheless, we can now provide an exemplar where translational optimality has not obviously selected on a set of codons distinct from the splice optimal set.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Simon Dittami for advice on expression resources. A.O.U. is a Royal Society Dorothy Hodgkin

Research Fellow and L.D.H. is a Royal Society Wolfsom Research Merit Award Holder. This work was supported by the a CONACyT scholarship (to J.M.T.-C.), the University of Bath (to X.W.), and the Erasmus program (to E.F.C.).

Literature Cited

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 52: 399–451.

Akashi H. 1994. Synonymous codon usage in *Drosophila* melanogaster: natural selection and translational accuracy. Genetics 136:927–935.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 25: 106–110.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST—database for "expressed sequence tags". Nat Genet. 4:332–333.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. Nat Genet. 30:29–30.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol. 62: 89–98.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 3:285–298.

Chamary JV, Hurst LD. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 21:256–259.

Chen L, Tovar-Corona JM, Urrutia AO. 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. Hum Mol Genet. 20:4422–4429.

Cock JM, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465:617–621.

Cock JM, et al. 2012. The *Ectocarpus* genome and brown algal genomics the Ectocarpus Genome Consortium. In: Piganeau G, editor. Genomic insights into the biology of algae. Amsterdam (The Netherlands): Elsevier. p. 141–184.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Eskesen ST, Eskesen FN, Ruvinsky A. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics 167:543–550.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2:549–555.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2:E268.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007–1013.

Fairbrother WG, et al. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. 32:W187–W190.

Franco P-Od, Rousvoal S, Tonon T, Boyen C. 2008. Whole genome survey of the glutathione transferase family in the brown algal model *Ectocarpus siliculosus*. Mar Genomics. 1:135–148.

Gilbert W, Marchionni M, McKnight G. 1986. On the antiquity of introns. Cell 46:151–154.

Goren A, et al. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. Mol Cell. 22:769–781.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. RNA 6:1197–1211.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends Genet. 23:321–325.

Ke S, et al. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 21:1360–1374.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35:125–131.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A. 108:11093–11098.

Logsdon JM. 1998. The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev. 8:637–648.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. Biochem J. 417:15–27.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 40:1413–1415.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23:301–309.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol Biol Evol. 24:1600–1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5: 343–353.

Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. 24:590–594.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141–1153.

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF. 1994. Testing the exon theory of genes—the evidence from protein-structure. Science 265:202–207.

Tanaka K, Watakabe A, Shimura Y. 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. Mol Cell Biol. 14: 1347–1354.

Taniguchi I, Masuyama K, Ohno M. 2007. Role of purine-rich exonic splicing enhancers in nuclear retention of pre-mRNAs. Proc Natl Acad Sci U S A. 104:13684–13689.

Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol Biol Evol. 24:2755–2762.

Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. Mol Syst Biol. 6:340.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol. 9:r29.

Webb CJ. 2005. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. Genes Dev. 19:242–254.

Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20:534–538.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875.

Zhang Z, et al. 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC Biol. 7:23.

Associate editor: John Archibald

Chapter III.

Why selection might be stronger when populations are small: intron size and density predicts within and between-species usage of exonic splice associated *cis*-motifs

Published manuscript:

XianMing Wu, Hurst LD. Why selection might be stronger when populations are small: intron size and density predicts within and between-species usage of exonic splice associated *cis*-motifs. ***Molecular Biology and Evolution*.** 2015 Jul;32(7):1847-61.

Contributions

All analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst.

Why Selection Might Be Stronger When Populations Are Small: Intron Size and Density Predict within and between-Species Usage of Exonic Splice Associated *cis*-Motifs

XianMing Wu[1] and Laurence D. Hurst*[,1]

[1]Department of Biology and Biochemistry, University of Bath, Bath, Somerset, United Kingdom

***Corresponding author:** E-mail: bssldh@bath.ac.uk.

Associate editor: Jianzhi Zhang

Abstract

The nearly neutral theory predicts that small effective population size provides the conditions for weakened selection. This is postulated to explain why our genome is more "bloated" than that of, for example, yeast, ours having large introns and large intergene spacer. If a bloated genome is also an error prone genome might it, however, be the case that selection for error-mitigating properties is stronger in our genome? We examine this notion using splicing as an exemplar, not least because large introns can predispose to noisy splicing. We thus ask whether, owing to genomic decay, selection for splice error-control mechanisms is stronger, not weaker, in species with large introns and small populations. In humans much information defining splice sites is in *cis*-exonic motifs, most notably exonic splice enhancers (ESEs). These act as splice-error control elements. Here then we ask whether within and between-species intron size is a predictor of the commonality of exonic *cis*-splicing motifs. We show that, as predicted, the proportion of synonymous sites that are ESE-associated and under selection in humans is weakly positively correlated with the size of the flanking intron. In a phylogenetically controlled framework, we observe, also as expected, that mean intron size is both predicted by $N_e\mu$ and is a good predictor of *cis*-motif usage across species, this usage coevolving with splice site definition. Unexpectedly, however, across taxa intron density is a better predictor of *cis*-motif usage than intron size. We propose that selection for splice-related motifs is driven by a need to avoid decoy splice sites that will be more common in genes with many and large introns. That intron number and density predict ESE usage within human genes is consistent with this, as is the finding of intragenic heterogeneity in ESE density. As intronic content and splice site usage across species is also well predicted by $N_e\mu$, the result also suggests an unusual circumstance in which selection (for *cis*-modifiers of splicing) might be stronger when population sizes are smaller, as here splicing is noisier, resulting in a greater need to control error-prone splicing.

Key words: **synonymous mutation, exonic splice enhancer, purifying selection, intron density.**

Introduction

Classical nearly neutral theory proposes that selection will be less efficient as the effective population size (N_e) goes down (Ohta 1973, 1992, 1996). In this context, we can, for example, interpret the finding that humans have a more "bloated" genome than seen in a species such as yeast which has a large effective population size and a correspondingly "lithe" genome (Lynch and Conery 2003). A lithe genome is one with short intergene spacer, relatively little repetitive sequence, few introns with the few found being relatively small. Might it, however, be the case that, as genomes decay owing to reduced N_e, the error rates of critical processes go up (cf. Frank 2007)? This might include increased mistranscription, mistranslation, missplicing, incorrect protein folding, incorrect phosphorylation, incorrect subcellular localization, etc. (Lynch 2007). Might this in turn then result in otherwise paradoxical stronger selection on error mitigation phenotypes when populations are small? Were this so, this would add a novel dimension to the nearly neutral hypothesis, as it would suggest that selection can sometimes be stronger when effective populations sizes are small, because, in this instance, the error rates are higher.

In the article, we examine this possibility by considering splicing error as an exemplar. In particular, we assume 1) that intron sizes tend to increase as N_e declines and that this is largely attributable to genome bloating (Lynch and Conery 2003) and 2) that within a genome exons flanked by larger introns have noisier splicing. As a consequence, we hypothesize that selection to reduce splice error rates will be more common in species with large introns, typically those with low N_e. Put differently, might humans have gradually expanded their introns through multiple small insertions, each being unable to be resisted by purifying selection, but in the process increased selection on modifiers of splicing in a ratchet-like process (cf. Frank 2007). The selection to reduce splice error rates we suggest will be manifested, in part, as a higher density of exonic *cis*-modifiers of splicing.

The two assumptions of our hypothesis appear to be reasonable, although the first of these has proven controversial. From phylogenetically uncontrolled correlation based

Article

analysis Lynch and Conery (2003) noted that across a wide span of species, as $N_e\mu$ declines introns tend to get larger and more common (higher density). $N_e\mu$ note is the product of effective population size (N_e) and the mutation rate (μ), the single statistic being estimated from population heterozygozity data. The trend in intron size Lynch and Conery attribute to weakening selection as N_e declines, that is, species with low N_e are less able to eliminate, through purifying selection, weakly deleterious insertion mutations when they occur in introns (and intergenic sequence). This study has, however, been criticized for failing to allow for phylogenic nonindependence between data points (Whitney and Garland 2010). Indeed, it was argued that the key result is not robust to proper phylogenetic control (Whitney and Garland 2010). As this $N_e\mu$ intron size/number correlation is a central tenet of the nearly neutral interpretation of genome anatomy, we return to this issue employing a phylogenetically controlled mode of analysis and more up to date estimates of $N_e\mu$, employing both more data and multiple modes of estimation. We show that with these updated estimates, in a phylogenetically controlled framework, $N_e\mu$ does indeed predict intron dimensions as Lynch and Conery (2003) postulated. We also show, however, that Whitney and Garland had an important objection, as we do not robustly recover this result using the original Lynch and Conery estimates of $N_e\mu$.

Our second supposition, that larger introns pose a threat to accurate splicing, has received experimental and comparative support. Notably, it is observed that experimental insertion of sequence into introns can reduce splice rates (Klinz and Gallwitz 1985; Luehrsen and Walbot 1992; Fox-Walsh et al. 2005; Sironen et al. 2006) and the exons hardest to splice consistently are those flanked by large introns (Bell et al. 1998; Fox-Walsh et al. 2005). Exons flanked by short introns, also associated with high expression levels, tend by contrast to be subject to less noisy splicing (Pickrell et al. 2010). In the longer term, exons flanked by long introns tend to be those most commonly lost (Kandul and Noor 2009), consistent with splice error rates being too high to sustain the exon. Exactly why exons flanked by larger introns are harder to splice is not fully understood, but one can speculate that if an intron is large, the splice site is harder to locate and the possibility for cryptic splice sites contained within the intron would be higher. The true splice sites need the reinforcement afforded by serine/arginine-rich (SR) proteins binding to exonic splice enhancers (ESEs).

Our hypothesis that selection to reduce splice error rates will be manifested in part as a higher density of exonic *cis*-modifiers of splicing is, in part, predicated upon the knowledge that *cis*-modifiers of splicing are known to be important in humans. For our genes, only approximately 50% of the information defining splice sites is at the splice site, the rest being in *cis*-motifs (Lim and Burge 2001). Possibly, the most importance of these motifs are ESEs (Blencowe 2000). The importance of ESEs is well demonstrated by the influence they have on selection on synonymous mutations (Carlini and Genut 2006; Parmley et al. 2006; Cáceres and Hurst 2013). Recent estimates suggest that around 4–5% of synonymous mutations in humans are under purifying selection because they disrupt ESEs (Cáceres and Hurst 2013). Our hypothesis might also predict that this figure might be a little lower in mice than in humans, as humans have on average larger introns. This has yet to be established, but suggestively, while standard nearly neutral N_e-based arguments would more obviously have predicted that selection on synonymous sites should be less common in humans than in rodents (Sharp et al. 1995; Keightley et al. 2005), the reverse seems to be true: An estimated 20% of synonymous mutations under selection in humans but only 10% in mice (Eory et al. 2010).

As prima facie support for the notion that selection for splice-error proofing can be more intense when populations are small, we note that the inferred centrality of ESEs to splicing in humans contrasts with species, such as yeast, with few/small introns and large populations. *Saccharomyces cerevisiae*, for example, appears not to employ ESEs to reinforce splicing (Spingola et al. 1999; Warnecke et al. 2008). More generally, the modes of selection on synonymous mutations in yeast and mammals appear to be rather different. Although in yeast there is easily identified translational selection (whereby codon usage evolves in accord with the tRNA pool), most acute in highly expressed genes (Ikemura 1982, 1985; Kanaya et al. 2001), the same is not robustly found in mammals (Bernardi et al. 1985; Sharp et al. 1995; Kanaya et al. 2001; Duret 2002). Rather, in mammals, selection on synonymous mutations is predominantly at exonic ends where ESEs aggregate (Carlini and Genut 2006; Parmley et al. 2006, 2007; Cáceres and Hurst 2013). In addition, however, there is evidence for selection on synonymous mutations in mammals mediated by miRNA pairing (Hurst 2006; Brest et al. 2011; Gartner et al. 2013), cotranslational folding (Lawrie et al. 2013), and mRNA structure modulation (Chamary and Hurst 2005; Nackley et al. 2006; Bartoszewski et al. 2010).

Our hypothesis makes a series of intra- and interspecific predictions. We expect, for example, that within a genome selection on ESEs might be more common in exons neighboring larger introns. Prior evidence supports the possibility that intron size is an important predictor of ESE density, at least within the human genome, ESEs being at a higher density at exon ends in proximity to longer introns (Dewey et al. 2006; Cáceres and Hurst 2013). It is not, however, known whether the higher density also implies more ESEs under selection. More generally, it is not known whether all putative ESE sites are functional. The apparent excess near long introns may, for example, reflect simple biased nucleotide content covarying with intron size (Duret et al. 1995). Here then we first ask whether selection on ESE-related synonymous sites might be more common in the vicinity of large introns, controlling for nucleotide usage. To this end we estimate the absolute number of ESE-related synonymous sites in proximity to an exon–intron junction that are under selection, as a function of the size of the flanking intron.

Our hypothesis also predicts that ESE usage should vary greatly between species, being greater when populations are small and introns large. Prior evidence suggests that there is

indeed considerable between-species variation in exonic *cis*-motif usage. Although ESEs are only well described in a handful of species, trends in *k*-mer usage across species in the vicinity of exon ends can be employed as a surrogate measure (Warnecke et al. 2008; Wu et al. 2013). Many *k*-mers are either enriched or depleted in the vicinity of exon junctions, trends in amino acid and codon usage in the vicinity of exon ends being a case in point. These trends are typically well predicted by underlying nucleotide content of the *k*-mers and the extent to which such nucleotides are employed in ESEs (Parmley and Hurst 2007; Cáceres and Hurst 2013), these being commonly purine-rich (Cáceres and Hurst 2013). Indeed, even in a species as distant from humans as *Ectocarpus* (a brown algae), 6-mer trends accord well with known human-described ESEs (Wu et al. 2013). Moreover, species lacking such distortion in *k*-mer usage also tend to be those that do not employ SR proteins to aid splicing, SR proteins being the binding partners of ESEs (Warnecke et al. 2008; Wu et al. 2013). Conversely, trends in *k*-mer usage in the vicinity of exon ends have been employed to define novel splice-related exonic motifs (Lim et al. 2011).

Taking the degree of distortion on *k*-mer usage in the vicinity of exon ends as a metric of the extent of *cis*-motif usage for splice control, prior studies report considerable variation between taxa in the number of *k*-mers affected (Warnecke et al. 2008; Wu et al. 2013). Here then, we ask whether we can account for this variation in terms of between-species variation in the size of introns and the effective population size. For compatibility with prior studies we employ in frame 3-mers, that is, codons. Prior evidence suggests that *cis*-motif usage, measured this way, may be most prevalent in species with more intronic sequence (Warnecke et al. 2008), but whether it is intron size or number that matters is not clear. Also suggestive of a relationship between ESE usage and intron dimensions, we recently showed that *Ectocarpus*, a species very distant from mammals and unusual in also having large introns, has extensive *cis*-motif usage, these motifs corresponding to ESEs well described in humans (Wu et al. 2013).

These prior analyses have, however, been confounded by a difference between species in the number of exons sampled and by not controlling for phylogeny. They also do not distinguish between intron density and intron size as predictors, whereas our model relates to intron size. We here ask in a phylogenetically explicit framework 1) whether mean intron size is a predictor of a species usage of *cis*-motifs and 2) whether it is a stronger predictor than intron density (the number of introns per bp of coding sequences [CDS]). To date, we are unaware of experimental evidence suggesting that intron density should predict ESE usage. This being so, if the selection across taxa for ESEs is mediated by changes in intron size alone, then intron density should not be a good predictor. In addition we employ a compound predictor, this being the ratio of CDS size to gene size that factors both intron density and mean intron size. If only intron size is relevant, then this compound predictor should be no better a predictor than mean intron size. Finally, we can ask how such trends in *cis*-motif usage correlate with N_e or rather $N_e\mu$, this metric estimated from intraspecific polymorphism levels. Given prior evidence that ESE usage and the nucleotides defining the splice site coevolve (Fairbrother et al. 2002; Dewey et al. 2006), we also address splice site usage as a function of $N_e\mu$ and intron size.

Results

Selection on Synonymous Mutations Is More Common When the Flanking Intron Is Large

The question as to whether selection on *cis*-splice motifs is more commonplace when the flanking intron is larger has two components: First, to what extent are such motifs under purifying selection as a function of the size of the flanking intron and second, how common are such motifs as a function of the size of the flanking intron.

Human ESEs Are Slower Evolving When the Flanking Intron Is Larger, but This Is Likely to Be Mutational

Are ESEs slower evolving than non-ESE sequence toward exon ends and are both the rate of evolution and the degree of constraint modulated by the size of the flanking intron? To address this, we consider human–macaque aligned sequence and classified exon ends in terms of the size of the flanking intron. As the exon ends are so small, to minimize estimation noise we consider for each intron size range the concatenation of all exon end alignments so as to provide a single estimate of Ks for each intron size class. We compare the synonymous rate of evolution in and out of ESE sequence.

To consider whether a hexamer might be an ESE motif, we took advantage of a recent analysis which derived two sets of motifs that were agreed on by the majority of ESE discovery analyses as being ESEs (Cáceres and Hurst 2013) and hence provide gold standard data sets with low false positive rates. As these are data sets that are intersects of independent data sets, in which at least three of four putative ESE data sets agree that a given hexamer motif is an ESE motif, we follow that prior nomenclature and refer to these as INT3 and INT3_400. Of the four original input data sets, one (Ke et al. 2011) presented a liberally defined set of ESEs and a more conservative top 400 set. As these two input sets are nonindependent, the prior authors (Cáceres and Hurst 2013) built two intersect data sets: One where the liberal set was employed and one in which the more conservative 400 strong data set was employed. The two resulting three-way intersect sets were thus termed INT3 (84 hexameric motifs with the liberal set employed) and INT3_400 (54 hexameric motifs with the top 400 hexamers employed), respectively.

As can be seen (fig. 1*A* and *B*) the ESE sequence evolves slower at synonymous sites than does non-ESE, as previously shown (Parmley et al. 2006; Cáceres and Hurst 2013). The difference between ESE and non-ESE may be a consequence of differences in mutation rate owing to skewed nucleotide content of ESEs. To examine this, we simulated sets of randomized pseudo-ESE sets that are the same size and drawn from the same underlying nucleotide content as the true ESE sets. We then match these pseudoESEs against the sequence alignments to determine the rate of evolution associated with

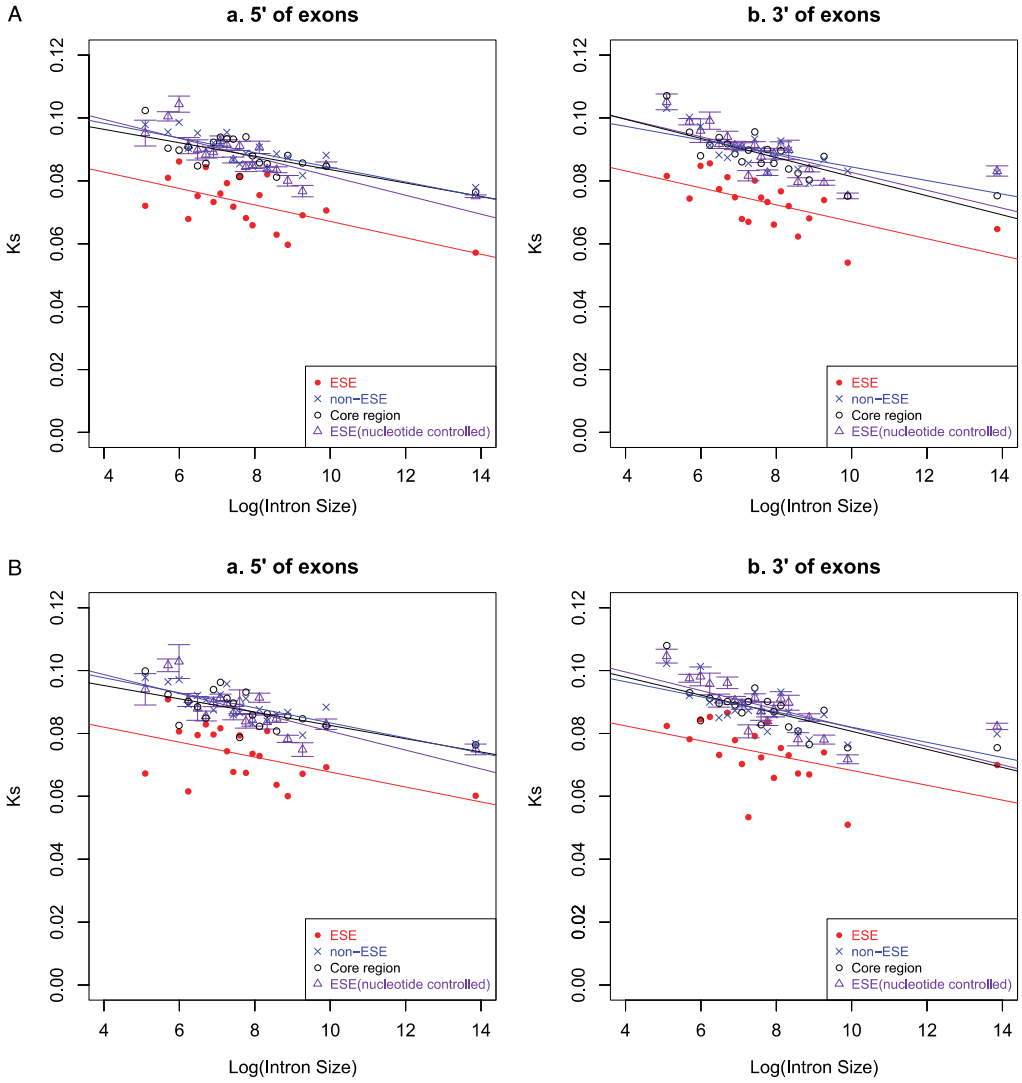


Fig. 1. Rate of synonymous evolution in ESE and non-ESE sequences at exon ends as a function of the Log of flanking intron size for two ESE data sets (*A*: INT3, *B*: INT3_400). In addition to Ks of ESE and non-ESE we also show Ks of exon core domains and pseudo-ESE, that is, hexamers of the same underlying nucleotide content as ESEs but not necessarily identified as being functional ESE. We consider 20 intron size bins apportioned so that all bins contain the same number of exon ends for concatenation, the numbers given reflecting the upper intron size limit of each bin.

these. These sets evolve faster than the true ESEs suggesting that ESEs are indeed (as commonly reported) under purifying selection (fig. 1), even allowing for biased nucleotide content.

More striking, we observe an evident negative correlation between Ks of ESEs at exon ends and the size of the flanking intron (INT3_400: 5′ rho = −0.50, $P = 0.027$; 3′ rho = −0.64, $P = 0.003$; INT3: 5′ rho = −0.53, $P = 0.018$; 3′ rho = −0.74, $P = 3 \times 10^{-4}$). This is consistent with stronger purifying selection on *cis*-splicing motifs or mutation rate differences covarying with intron size. That we see a commensurate decrease in Ks of the "non-ESE" sequence as a function of intron size might reflect either 1) purifying selection in exon ends is generally stronger in the vicinity of large introns, possibly because the definition of non-ESE is too liberal and includes much sequence that is functional splice related motif or 2) the mutation rate in exons is lower in the vicinity of larger introns. To examine the latter possibility we compare Ks of exon cores as a function of the size of neighboring introns (we consider the size of the 5′ and 3′ intron separately), under the presumption that little or no sequence in exon cores will modulate splicing. We observe that Ks of cores also show a decreasing tendency as intron sizes increases (fig. 1). Although

we can conclude that the reduced rate of evolution of ESEs, compared with non-ESE and pseudo-ESE in the same exons, is not solely mutational in origin, we cannot then exclude the possibility that the low rate of synonymous evolution at exon ends in the vicinity of large intron is at least in part owing to genomically regional mutation rate biases in the vicinity of large introns.

Consideration of the rate of synonymous evolution of exon cores also permits us to define the approximate degree of constraint operating on ESE at exon ends as:

$$\text{Flank ESE constraint} = \frac{[\text{Ks core} - \text{Ks ESE flank}]}{\text{Ks core}}.$$

This may be conservative, but it is noteworthy that Ks non-ESE flank, Ks pseudoESE, and Ks core are all approximately of the same magnitude (fig. 1*A* and *B*), much higher than Ks ESE flank. In the absence of purifying selection on ESE at exon flanks, in excess of that at exonic cores, the degree of constraint should be zero. We observe that the level of constraint, thus defined, operating on ESEs at exon flanks is not significantly related to the size of the flanking intron, although Spearman's rho is positive in all incidences (fig. 2*A* and *B*; supplementary table S1.1). What can reasonably be concluded is that selection on ESEs is not obviously weaker in the vicinity of large introns. To estimate the number of sites under selection at exon flanks, we need in addition to factor in not just the level of constraint but also ESE density. This we consider next.

Allowing for Increased ESE Density in Proximity to Large Introns, Selection on Synonymous Sites Associated with ESEs Is (slightly) More Common When Introns Are Larger

It has previously been reported that ESE density tends to be a little higher in the vicinity of longer introns (Dewey et al. 2006; Cáceres and Hurst 2013). We replicate this by partial correlation analysis between ESE density and three intronic dimensions (supplementary table S2.1). For ESE data set INT3, both 5′ and 3′ show significant correlation between ESE density and mean intron size (5′ rho = 0.03, $P = 9 \times 10^{-4}$; 3′ rho = 0.03, $P = 9 \times 10^{-4}$). However for the smaller INT3_400 ESE data set, 3′ correlation is not significant (INT3_400: 5′ rho = 0.06, $P = 2 \times 10^{-13}$; 3′ rho = −0.01, $P = 0.14$). More marginal results at exonic 3′-ends is a common theme in our analyses which we comment on later.

To evaluate the net effect of flanking intron size (constraint and increased density), we calculate the proportion of synonymous sites under ESE-related constraint at exon flanks as flank ESE constraint × ESE density. It is no surprise that the net effect of flanking intron size on proportion of sites under selection is an increasing function, albeit only weakly so, as both underlying trends are positive. However, using the conservative binning method ($N = 20$) the trend is not significant. This may well reflect a limited sample size ($N = 20$). To avoid this problem, we instead calculate the regression line of logarithm value of flank intron size versus ESE constraint (using unbinned data). Using this regression line we then estimate the mean ESE constraint for exon flanks given the size of the neighbor intron. For each exon individually, we then calculate ESE density × regression estimated constraint. We find in all cases a positive and highly significant Spearman's rank correlation (INT3_400: 5′ rho = 0.439, $P = 0$; 3′ rho = 0.123, $P = 2.7 \times 10^{-31}$; INT3: 5′ rho = 0.070, $P = 2.5 \times 10^{-13}$; 3′ rho = 0.646, $P = 0$).

Some of these latter values appear unusually high, which may relate to our interpolation method which smoothes out noise resulting from the tiny number of sites contributing to constraint estimates at individual exon ends. Moreover, this method does not allow for potential covariance with intron number and intron density. To examine this, we analyze gene level (rather than individual exon end) metrics. For each gene, we consider the intron density, intron number, and mean intron size. In addition, we consider the constraint revealed in concatenated exon flanks and concatenated exon cores. A small minority of genes have no synonymous site evolution in exon flank ESEs giving a constraint of unity. As Spearman's correlation is not necessarily robust to tied values, we thus also replicated analyses using Goodman–Kruskall gamma test with *P* determined by simulation (supplementary table S3.1). To minimize estimation noise, we require for all genes a minimum of 102 bp of concatenated sequence. We then ask whether mean intron size is related to constraint on ESE at flanks.

We find that mean intron size is positively and significantly correlated with 5′ but not 3′ ESE constraint (supplementary table S3.1). This trend is weak (rho 0.048–0.056) but is reported using both statistics and both ESE data sets. Intron density and intron number are not significant predictors. Partial spearman correlation also reports mean intron size as a significant predictor (supplementary table S3.2). To some degree, these results are not robust to increasing the minimum threshold length for analysis from 102 to 150 bp or higher. This, however, appears to be a consequence of reduced sample size. We resampled genes by the number of 150-bp cutoff group from the gene pool of 102-cutoff group and repeated 1,000 times to find how often the intron dimensions can significantly predict the constraints. For mean intron size, only in a little over than 60% of resamplings can we still see significant Spearman partial correlation with 5′-ends ESE constraint for both data sets. For Goodman and Kruskal's gamma, the commensurate figure is around 42% (supplementary table S3.3). This accords with the trends being weak and hence sensitive to sample size reductions.

We conclude that in the human genome mutations in ESEs at exon ends are probably more commonly under selection when the flanking intron is larger, the effect being mostly mediated by an increased ESE density. This result in turn suggests that disease-associated mutations might be slightly more common in exon ends in the vicinity of large introns, but the effect appears to be modest.

We note that as regards this result we are agnostic as to the cause. This may be a direct effect of intron size or owing to a covariance between intron size and splice site strength, possibly with expression level as a covariate. Our intention here is not to distinguish between these explanations, but simply to suppose that this evidence provides prima

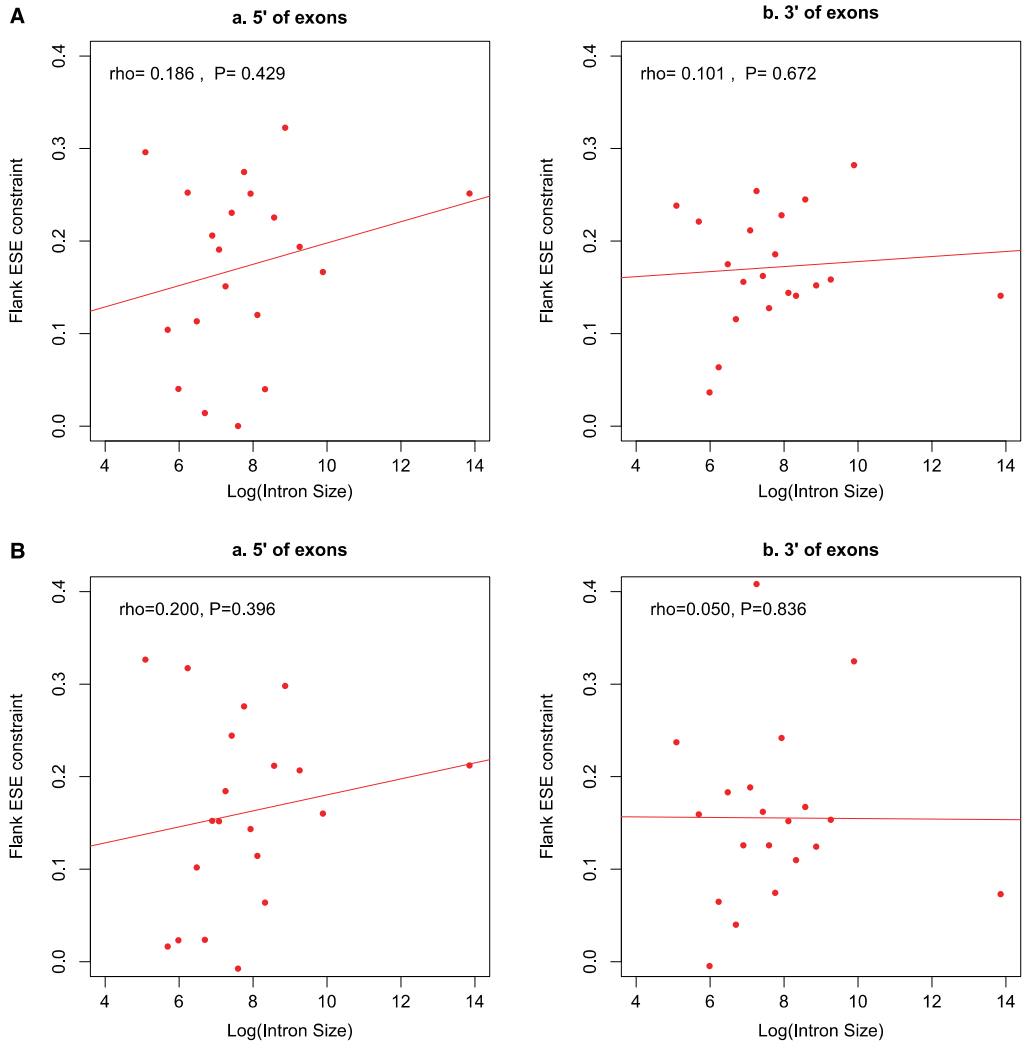


FIG. 2. The degree of selective constraint on ESE sequences at exon ends as a function of the Log of flanking intron size for two ESE data sets (*A*: INT3, *B*: INT3_400). For definition of constraint, see main text. For intron size definition, see figure 1. Note that in all cases constraint appears stronger when intron sizes are larger, although using 20 bins the trends are not significant.

facie support to the hypothesis that an increase in intronic dimensions within a species can be coupled with more selection for *cis*-modifiers of splicing. We note, however, that a model that supposes that ESEs are used more in exons next to long introns might be a means to increase elongation rate, to compensate for the time to process the longer intron, is not well supported (see supplementary tables S4.1–S4.2).

The Ratio of Mature CDS to Gene Size Is the Best Predictor of between-Species *cis*-Motif Usage

Given the above result and prior experimental and comparative data on the difficulty of splicing exons when the neighboring intron is large (see Introduction), we might expect that mean intron size would be a predictor of the commonality of the usage of exon flank *cis*-modifiers of splicing. To establish the latter we consider, for 30 highly phylogenetically dispersed species, the proportion of codons or amino acids that show significant trends in their usage as a function of the distance from an exon–intron junction, these metrics having been shown previously to correspond well with ESE motif usage (Parmley and Hurst 2007; Parmley et al. 2007; Warnecke et al. 2008; Cáceres and Hurst 2013). Using a Bayesian comparative framework, we can then ask whether mean intron size is indeed a good predictor of *cis*-motif usage. We find that it is (table 1). Although our

Table 1. Evidence for Phylogenetically Controlled Correlation between Amino Acid/Codon Usage Trends and the Genomic Traits.

	All Exons (AA)	All Exons (codon)	Random 5,000 Exons (AA)	Random 5,000 Exons (codon)
Log BF ($Y \sim X$)[a]	48.241	39.394	31.923	42.027
Log BF ($Y \sim N$)	37.484	29.202	24.055	32.294
Log BF ($Y \sim M$)	20.145	15.018	12.214	18.410

NOTE.—We employ two metrics of skews at exon ends, the number of codons showing a significant skew and the number of amino acids showing a significant skew. For each, in addition we report results wherein for each species all relevant exons are employed and a second metric where the input sample size is the same for all species (5,000 randomly chosen exons). In the latter instance, we consider the mean number of significant trends from multiple samplings of 5,000 randomly chosen exons. *Y*, proportion of amino acids/codons showing significant trends; *X*, mean CDS length/gene length; *N*, introns per kb exon; *M*, mean intron size.

[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]), is the test statistic of BayesTraits which gives the information of evidence for correlated evolution: Weak evidence (<2), positive evidence (> 2), strong evidence (5–10), very strong evidence (> 10). All Log BF values in the table are greater than 10, so the evidence from all correlations is very strong.

measure of the degree of this trend controlling for the size of exonic input data set is the preferred metric, we show that usage of an uncontrolled metric (employing all valid exons within a species) does not distort the picture (table 1). Hereafter we employ the sample size controlled metric exclusively, unless mentioned otherwise.

As a control test, we consider intron density. Intron density, measured as number of introns per kilobase of mature CDS, holds no information regarding the size of the introns and hence if size is the key variable density should be irrelevant. Unexpectedly, not only do we find that density is a predictor of the extent of *cis*-motif usage, we also observe that it is consistently a better predictor than intron size (the BayesTrait score is higher in all modes of analysis; table 1). Given this surprising result we ask whether a metric that considers the net effect of density and size might be an even better predictor. To this end, we employ the ratio of mature CDS to gene size (alias immature transcript size). This is consistently the best predictor (table 1). We conclude that intron size alone is not adequate to describe the between-species trends in *cis*-motif usage and that density effects are also of relevance. The logic of the importance of the density effects we discuss below.

Evidence for Coevolution of Splice Site and *cis*-Motif Usage

Prior evidence suggests that ESE usage is higher in proximity to certain splice sites (Berget 1995; Graveley 2000). One possibility is that "weak" splice sites might be more in need of the reinforcement offered by flanking ESEs (Fairbrother et al. 2002). In support of this, ESE density appears to be stronger in proximity to "weak" splice sites (Dewey et al. 2006; Plass et al. 2008; Cáceres and Hurst 2013). To ask whether ESE usage across species was predicted by relative usage of different splice sites, we investigated all splice sites across 30 species. The splice sites we represented as four-letter nucleotide strings, nucleotides of exons in upper case, nucleotides of introns in lower case. After phylogenetic correction, BayesTraits provided very strong evidence for correlation between usage of *cis*-motif and usage of two splice sites ("AGgt" and "agGT") (table 2). This indicates a preference of exonic splice associated *cis*-motifs to these specific splice sites. These results indicate that the trends in *cis*-motif usage across species reflect in part coevolution with splice site usage.

$N_e\mu$ Predicts Intronic Dimensions

Given that intronic dimensions predict *cis*-motif usage across taxa, what, we can ask, predicts intronic dimensions across taxa? An attractive proposal is that introns and intronic sequence accumulate owing to weakened selection against insertions associated with reduced N_e. Previously, Lynch and Conery (2003) have argued, in a phylogenetically uncontrolled analysis, that intronic size can be well understood in the context of such a nearly neutral model. They posit that as N_e reduces selection becomes weaker and the ability of a species to resist weakly deleterious insertions (both new introns and new sequence within extant introns) is in turn reduced. Thus, they predict large introns and high density of introns in species with low N_e.

Their analysis has been criticized on numerous fronts, not least of which is the assumption of $N_e\mu$ is a good predictor of the behavior of N_e alone (Daubin and Moran 2004) (a problem our analysis is also sensitive to). Further, they estimated $N_e\mu$ for a sample of species often employing limited sequence data. Perhaps most importantly, their analysis was criticized for failing to control for phylogenetic structure, in effect assuming a star phylogeny (Whitney and Garland 2010). This same follow-up analysis, employing a phylogenetically explicit method failed to observe a relationship between genome size parameters and N_e. We return to this issue employing three methods to estimate $N_e\mu$, three metrics of intronic content, and a fully controlled phylogenetic methodology.

Three $N_e\mu$ values of this study show very significant correlations between themselves; however, our estimates of $N_e\mu$ do not correlate well with those of Lynch and Conery (table 3, supplementary fig. S1, the blue line indicates the standard major axis [SMA] regression). We find that our $N_e\mu$ estimates robustly predict all three intronic dimensions in the expected direction (table 4). By contrast, we can replicate Whitney and Garland's failure to detect such a correspondence: After phylogenetic correction, although there is a strong evidence to support the correlation between $N_e\mu$ values of Lynch and Conery's study and the ratio of mature CDS to gene size, these $N_e\mu$ values do not correlate well with intron density and mean intron size (table 4). We suggest that the paucity of data contributing to the Lynch and Conery estimates of $N_e\mu$ is the major issue with their analysis.

Table 2. *Cis*-Motif Usage Correlates Significantly with Usage of "AGgt" and "agGT" Splice Sites.

	All Exons (AA)	All Exons (codon)	Random 5,000 Exons (AA)	Random 5,000 Exons (codon)
Log BF ($Y \sim P1$)	39.0359	31.7091	26.4632	33.7518
Log BF[a] ($Y \sim P2$)	52.1594	64.0366	76.8355	58.1153

NOTE.—*Y*, proportion of amino acids/codons showing significant trends; *P1*, proportion of AGgt (Capital letter: exon, small letter: intron); *P2*, proportion of agGT.
[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]). All Log BF values in the table are greater than 10, so the evidences of all correlations (positive) are very strong.

Table 3. Spearman's Correlation Analysis Results for $N_e\mu$ Values of This Study and the Prior Study of Lynch and Conery.

	rho	rho^2	*P*
$N_e\mu$_Eta ~ $N_e\mu$_Lynch	0.093	0.009	0.765
$N_e\mu$_Pi ~ $N_e\mu$_Lynch	0.165	0.027	0.591
$N_e\mu$_S ~ $N_e\mu$_Lynch	0.093	0.009	0.765
$N_e\mu$_Eta ~ $N_e\mu$_S	0.996	0.991	0.000
$N_e\mu$_Pi ~ $N_e\mu$_Eta	0.970	0.941	0.000
$N_e\mu$_Pi ~ $N_e\mu$_S	0.975	0.951	0.000

NOTE.—We compare our three different estimators for $N_e\mu$, (Eta, Pi, and S) and Lynch's single estimate.

$N_e\mu$ Predicts Splice Site Usage but Not *cis*-Motif Usage

The above sets of results suggest a simple narrative to explain *cis*-motif usage across species. As N_e declines, so introns become more abundant and larger, owing to the weakening of purifying selection (result 4 above). A consequence of this is that small insertions may accumulate in a ratchet-like manner. Similarly, splice sites might decay. Both splice site decay and the increase in intron size cause increases in the rate of missplicing compensated by increased usage of exonic *cis*-motifs. Within genomes, the argument goes, this is reflected in a higher density of functional *cis*-motifs in the flanks of exons that neighbor large introns (result 1) and associated with particular splice sites (Cáceres and Hurst 2013). Thus, selection on synonymous mutations at exon flanks is more common when the flanking intron is large (result 1 above) and species with on average larger introns have more *cis*-modifiers (result 2), these being especially common when certain splice sites become more common (result 3). Additionally, consistent with ESE-splice site coevolution, we see intraspecifically that AGgt exons are flanked by larger introns (supplementary table S5.1), consistent with splice site—ESE - intron size three-way coevolution. We would thus expect that $N_e\mu$ should also in turn predict the usage of *cis*-splice modifiers and splice sites.

The latter result we find to be robustly supported, at least for 5′-end splice site usage. More specifically, the correlations between $N_e\mu$ values and the usage of "AGgt" (i.e., 5′-splice site) are very strong, whereas those about the usage of "agGT" (3′-splice site) are weak (table 5).

Do we also find that $N_e\mu$ predicts *cis*-motif usage? This result we have yet to demonstrate. The prediction we make is that species with low $N_e\mu$ will be species with more common skews in codon or amino acid usage owing to selection for *cis*-modifiers of splicing. Unexpectedly, despite having observed all prior correlations ($N_e\mu$ predicts intron dimensions and splice site usage, intron dimensions and splice site usage predict *cis*-motif usage), we fail to recover a trend whereby *cis*-motif usage is predicted by $N_e\mu$ (table 6). For the $N_e\mu$ estimator *S*, there may be a weak trend but for others there is no evidence. Employing the sample size uncorrected measure of the number of trends removes any weak trend reported for *S* (table 6). We conclude that we find evidence that splice site usage, but not *cis*-motif usage, correlates with $N_e\mu$.

Alternative Splicing Rate Does Not Explain *cis*-Motif Usage

One reason that $N_e\mu$ might not predict *cis*-motif usage is that other covariates are important and mask any effect. A potentially key covariable might be the frequency of alternative splicing. We observed previously that the brown algae *Ectocarpus* has a striking number of codons and amino acids showing skews in usage in the vicinity of exon junctions, many more indeed than humans (Wu et al. 2013). This we hypothesized may reflect the low rate of alternative splicing that we could detect. If alternative splicing is rare in a species, then more of the annotated exons will be under selection to be properly spliced more of the time. Alternatively, ESEs might modulate alternative splicing, which is more common in "complex" species (Chen et al. 2014), typically with low N_e. Note that these two models make opposite predictions.

To provide an assessment of this, we consider transcript depth-controlled estimates of the rate of alternative splicing for 14 species (Chen et al. 2014). We find strong evidence to support the correlation between alternative splicing rates with the ratio of mature CDS to gene size. Although intron density is a better predictor of *cis*-motifs than is intron size, the correlation between alternative splicing rates and mean intron size is better than that with intron density (table 7). Between-species differences in alternative splicing rates do not, however, predict between-species trends in *cis*-motif usage very well (supplementary table S5.2). We conclude that although alternative splicing rates and intronic dimensions covary, the former appears not to explain trends in *cis*-motif usage.

Is the Commonality of Decoy Splice Sites the Main Driver of Splice Associated *cis*-Motif Usage?

Why might it be that the best between-species predictor of *cis*-motif usage was not simply mean intron size, but an

Table 4. Evidence for Phylogenetically Controlled Correlation between $N_e\mu$ Values and Splice-Related Genomic Traits.

	X	N	M
Log BF ($N_e\mu$_Pi ~ Splice-related Genomic Traits)[a]	15.762	23.424	41.057
Log BF ($N_e\mu$_S ~ Splice-related Genomic Traits)	14.572	22.590	39.944
Log BF ($N_e\mu$_Eta ~ Splice-related Genomic Traits)	13.988	22.695	40.367
Log BF ($N_e\mu$_Lynch[b] ~ Splice-related Genomic Traits)	5.290	0.989	−0.587

NOTE.—We employ our three different estimators for $N_e\mu$ (Eta, Pi, and S) and Lynch's single estimate. *X*, mean CDS length/gene length; *N*, introns per kb exon; *M*, mean intron size.

[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]), is the test statistic of BayesTraits which gives the information of evidence for correlated evolution: weak evidence (<2), positive evidence (> 2), strong evidence (5–10), very strong evidence (> 10). All Log BF values in the table are greater than 10, so the evidences of all correlations are very strong.

[b]This $N_e\mu$ value is from previous study (Lynch and Conery 2003).

Table 5. Evidence for Phylogenetically Controlled Correlations between $N_e\mu$ Values and Usage of "AGgt" (very strong) and "agGT" (weak) Splice Sites Using Three Estimators of $N_e\mu$, Namely Pi, S, and Eta.

	$N_e\mu$_Pi	$N_e\mu$_S	$N_e\mu$_Eta
Log BF ($N_e\mu$[a] ~ *P1*)	22.7225	19.1016	20.6161
Log BF ($N_e\mu$ ~ *P2*)	1.6456	−0.1762	0.6543

NOTE.—*P1*, proportion of AGgt (Capital letter: exon, small letter: intron); *P2*, proportion of agGT; Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]).

[a]Three types of $N_e\mu$ ($N_e\mu$_Pi, $N_e\mu$_S, $N_e\mu$_Eta).

aggregate measure of size and density? From the logic that we laid out (difficulty of exon junction recognition in the context a large flanking intron), this is perhaps unexpected. We suggest that the problem may be one of decoy splice sites. Imagine a gene with one large intron and no residues elsewhere downstream of the true 3′ splice site that might be recognized as a possible acceptor site. Would such a gene have error-prone splicing? We would suggest not if there is a unique strong site (the true acceptor site) compatible with splicing. By contrast, by definition, the same gene with two introns must have at least two putative acceptor sites. Thus the more introns and the weaker the splice sites, the more potential there is for missplicing.

This suggests then a simple explanation for why intron density matters. We assume that SR proteins bound to the immature RNA accumulate at exon ends bound to ESEs. A given 5′ splice site, we assume also, tends to attach to the perceived nearest 3′ splice site, this being identified by the accumulation of ESEs and SR proteins. The extent of accumulation of ESEs we suggest is a function of the chance the splice site might be "missed." Strong splice sites in close proximity (short introns) are unlikely to be missed and hence need little reinforcement. By contrast ESEs are needed more in the vicinity of larger introns as the ability to find the nearest 3′ splice site is harder owing to the distance and because the number of decoy sites is higher, that is, when the density of introns is higher. However, whether it is density per se or absolute number of introns that is key is not immediately transparent, as it is unclear whether the absolute proximity of decoy splice sites to the "real" splice site is relevant. If physical proximity is relevant then density may matter, if not absolute number may be more important.

Such a model makes an intragenomic prediction, namely that controlling for intron length, intron density or number should predict ESE density. From the partial correlation analysis between ESE density and three intronic dimensions (mean intron size, intron density, and intron number), all 5′-end calculations show very significant partial correlations, regardless of the choice of ESE data set. At the 3′-end, the result is less clear. For INT3 ESE data set, the correlation between ESE density and intron density is not significant, whereas intron number and intron size are predictors. At 3′−ends, all partial correlations for INT3_400 are not significant (supplementary table S2.1). The correlation with intron number is perhaps the most revealing, suggesting that density per se functions as a proxy to absolute number and hence that exon size considerations are not so relevant. Further, these results suggest that, although ESE usage at 5′- and 3′-ends of exons is usually considered to be symmetrical in humans (motifs commonly found at 5′-ends tend to be common at 3′-ends [Warnecke et al. 2008; Lim et al. 2011]), that at least as regards intron density mediated effects 5′- and 3′-ends are under different modes of selection. The suggestive evidence that net selection on ESEs is better correlated with intron length for 5′- ends than 3′-ends supports the same proposition, as does the 5′−3′ difference in splice site predicted by $N_e\mu$.

If the problem faced is one in which downstream exons and introns presenting decoy splice sites, then we might also expect a difference in ESE density within a gene, as different exons have a different number of downstream introns and exons and hence a different number of potential decoy splice sites. We address this by comparing the ESE density at the 5′-end of the second exon in a gene and the 5′-ESE density at the last but one exon in genes with at least four exons. We do not employ the very last exon owing to possible constraints on nucleotide content in the vicinity of the stop codon.

We find strong evidence that intragene location matters, with ESE density higher earlier in a gene. From comparing the ESE density at the 5′-end of the second exon in a gene and the 5′-ESE density at the last but one exon in genes with at least four exons, the medians of ESE density of last but one exons and second exons are about 2-fold different (INT3 ESE density: 0.0909 and 0.1739, INT3_400 ESE density: 0.0882 and 0.1739), in last but one and second exon, respectively (supplementary table S6.1). To examine the significance of this we perform a paired test, comparing the ESE density within the same gene between the two exon 5′-flanks. Results are as expected of the decoy splice site model. For INT3 data set, the number of genes which show ESE density of the second exon to be higher than that of last but one exon, reaches 491 and the number where ESE density of second exon is relatively lower is 381 (binomial test $P = 2.6 \times 10^{-5}$). For INT3_400 data set, the corresponding values are 332 (ESE density of second

Table 6. Little Evidence for a Phylogenetically Controlled Correlation between $N_e\mu$ Values and Amino Acid/Codon Usage Trends (Y).

	All Exons (AA)	All Exons (codon)	Random 5,000 Exons (AA)	Random 5,000 Exons (codon)
Log BF ($N_e\mu$_Pi ~ Y)[a]	−0.486	−2.065	−2.693	−4.436
Log BF ($N_e\mu$_S ~ Y)	−1.383	−0.206	1.514	0.728
Log BF ($N_e\mu$_Eta ~ Y)	0.534	−0.520	−2.038	1.079

NOTE.—We employ our three different estimators for $N_e\mu$ (Eta, Pi, and S) and four metrics of *k*-mer usage. *Y*, proportion of amino acids/codons showing significant trends.
[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]), is the test statistic of BayesTraits which gives the information of evidence for correlated evolution: weak evidence (<2), positive evidence (> 2), strong evidence (5–10), very strong evidence (> 10). All Log BF values in the table are less than 2, so the evidences of all correlations are weak.

Table 7. Evidence for Correlation between Alternative Splicing Rates and Splice-Related Genomic Traits.

	X	N	M
Log BF (ASL1 ~ Splice-related Genomic Traits)[a]	5.259	2.782	7.299
Log BF (ASL2 ~ Splice-related Genomic Traits)	8.714	4.589	9.500

ASL1, average number of ASEs per gene (residual of the polynomial regression between num of ESTs [col. O] and ASL [col. U]); ASL2, average number of ASEs per gene (residual of the linear regression between the log-transformed num of ESTs [col. O] and ASL [col. U]); *X*, mean CDS length/gene length; *N*, introns per kb exon; *M*, mean intron size.
[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)] − log [harmonic mean (simple model)]), is the test statistic of BayesTraits which gives the information of evidence for correlated evolution: Weak evidence (<2), positive evidence (> 2), strong evidence (5–10), very strong evidence (> 10).

exon is higher) and 240 (ESE density of second exon is lower), again supporting a higher density in second exons (binomial test $P = 2.0 \times 10^{-5}$).

To test whether the trend is owing to confounding effects of proximal intron size, we employed Mann–Whitney *U* test to analyze residuals of a loess regression while 5′ proximal intron size is being controlled. Results are again as predicted by the decoy model. We again find a significantly higher ESE density at 5′ end of second exons compared with last but one exons (Mann–Whitney *U* test comparing residuals of 5′ intron size vs. ESE density, INT3: $P = 2.42 \times 10^{-71}$, INT3_400: $P = 5.32 \times 10^{-46}$). A within-gene paired test on residuals from above loess regression supports the same conclusions (INT3_400: number of higher second exon residuals = 386, number of lower second exon residuals = 289, binomial test $P = 2.85 \times 10^{-5}$; INT3: number of higher second exon residuals = 566, number of lower second exon residuals = 404, $P = 3.25 \times 10^{-8}$). A higher density of ESEs earlier in a gene is, we suggest, consistent with the decoy model given that early exons by definition have more downstream splice sites than do later ones. It also suggests a novel (to our knowledge) model of splicing reinforcement that is different in different sections of the same gene.

Discussion

We conjectured that reduced N_e might lead to larger introns and weakened splice sites, which in turn could lead to stronger selection for motifs that keep in check the increase in the degree of error-prone splicing. All results bar one support this. We find that synonymous sites are more commonly under selection within humans when exons are flanked by larger introns (largely because more sites function as *cis*-motifs), that intronic dimensions and splice site usage predict *cis*-motif usage across species, and that $N_e\mu$ predicts intronic dimensions and splice site usage (we note that this tidies up the prior objection that in a phylogenetic framework the results of Lynch and Conery do not hold [Whitney and Garland 2010]). In addition, we find that intraspecifically, exons flanked by large introns both have higher ESE density and greater usage of AGgt, consistent with coevolution between splice site, ESEs and intron size. What we do not observe is that $N_e\mu$ predicts *cis*-motif usage.

Given the support for the hypothesis from all but one of the tests, we suggest that it would be premature to reject the hypothesis out of hand. Indeed, one possibility is that our estimation of $N_e\mu$ is either too rough or otherwise flawed. It is striking, for example, that our estimation and that of Lynch and Conery do not correlate well, despite being based on the same underlying premise. Moreover there might be a systematic issue with all polymorphism-based attempts to estimate $N_e\mu$, this being that the expected correlation between N_e and heterozygozity appears to be much weaker than predicted by the neutral model (which forms the basis for $N_e\mu$ estimation). Gillespie (2001) argues that the approximate invariance (or weak positive correlation) between N_e and heterozygozity is owing to an increased rate of positive selection when populations are large, thereby causing regular collapses of heterozygozity owing to hitchhiking type effects. We do not wish to comment on the veracity of this claim, but simply wish to note that of all the variables that we have employed, $N_e\mu$ is the one we have least confidence in, both as regards its estimation and its interpretation. Recent evidence that intraspecific diversity is predicted by life-history traits (Romiguier et al. 2014) adds to the notion that a relationship between $N_e\mu$, deduced from heterozygozity data, and the strength of selection may be compounded by covariates. Nonetheless, we observe that $N_e\mu$ robustly predicts intronic dimensions and splice site usage, suggesting that it is perhaps not too poor an estimator.

Although we have framed the above hypotheses and results in the context of the nearly neutral model, the same results might, however, also be consistent with a model in which increasing *cis*-motif usage across taxa reflects greater tissue or cell type diversity, ESEs then operating as providers of tissue-specific alternative splice patterns. It is indeed observed that species with more cell types do have more alternative transcripts (Chen et al. 2014). Might this coupling be explained by increased usage of ESEs? Our and other results suggest not. We observe no relationship between *cis*-motif usage and alternative splicing rates. Moreover, there is no

strong prior evidence to suppose that ESE usage is a modulator of alternative splicing. Indeed, although our intersect data sets find no difference in ESE density between alternative and constitutive exons (Cáceres and Hurst 2013), an experimentally defined set of exonic splice modifiers (Ke et al. 2011) found a much higher ESE density in constitutive than in alternative exons. Earlier reports also indicated that, although conserved alternative exons have very low rates of evolution, this was not owing to especially strong constraint on ESEs (Parmley et al. 2006; Cáceres and Hurst 2013). These results thus suggest that ESEs are not there as elements to control alternative splicing forms, but rather to make more robust the splicing of constitutive exons, especially those with weak splice sites. For these reasons, we suggest that higher transcript diversity in species with small population sizes/multiple tissue types is not an easily defendable explanation for the trends in *cis*-motif usage.

An unexpected result was that in the between-species comparison, intron size is by no means the best intron-dimension predictor of *cis*-motif usage. Rather a combination of size and density is a much better predictor. We propose a decoy splice spite model as a potential explanation. This model correctly predicts intragenomic and intragenic trends, highlighting the selection on the earliest exons as being especially acute. The intragenic trend may however have an alternative explanation, namely that it is simply more damaging to missplice an early exon than it is to missplice a later exon. For example, the downstream effects of a frame-shifting splice event may be different for the two. It is not so obvious that such an argument can explain the intragenomic, intergenic trends (i.e., mean intron size, intron density, and intron number all independently predict 5′-ESE usage). This model and the apparent asymmetry between 5′- and 3′-effects are, we suggest, worthy of further scrutiny.

Materials and Methods

Exon and Intron Sequences from 30 Species

From "Table Browser" of UCSC (http://genome.ucsc.edu/cgi-bin/hgTables, last accessed January 23, 2014) and FTP site of NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes, last accessed January 23, 2014), we obtained all available genes from 30 species (*Anolis carolinensis, Anopheles gambiae, Arabidopsis thaliana, Brachypodium distachyon, Caenorhabditis elegans, Callithrix jacchus, Cryptococcus neoformans, Dictyostelium discoideum, Drosophila melanogaster, Danio rerio, Ectocarpus siliculosus, Gallus gallus, Gorilla gorilla, Homo sapiens, Ictidomys tridecemlineatus, Meleagris gallopavo, Macaca mulatta, Mus musculus, Oryzias latipes, Oryza sativa, Pongo abelii, Plasmodium falciparum, Paramecium tetraurelia, Pan troglodytes, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Strongylocentrotus purpuratus, Sus scrofa, Takifugu rubripes, Xenopus tropicalis*). Sequences without the normal start (ATG) and stop codons (TAA, TAG, and TGA), with internal stop codons, ambiguous nucleotides ("N"), and without introns were all removed from the data set (supplementary table S7.1).

Determining Trends in Amino Acid and Codon Usage

In previous analyses, codons preferred near exon ends were well predicted by the composition of experimentally defined ESEs (Parmley and Hurst 2007; Cáceres and Hurst 2013). We thus presume that the frequency of distorted codon or amino acid usage in vicinity of exon junctions is a fair measure of *cis*-splice motif usage. The trend in usage of each codon and amino acid was investigated as a function of the distance from the exon–intron boundary up to a distance of 34 codons (to accord with an earlier analysis [Warnecke et al. 2008]). The 5′- and 3′-ends were analyzed separately with the codon in direct proximity to the boundary being eliminated and the first and last exons being excluded. For each codon and amino acid under consideration, we determined, after Bonferroni correction, rho and *P* value by two-tailed Spearman correlation of proportional usage as a function of distance from the boundary. A negative rho indicates a codon or amino acid that is preferred near exon ends, whereas a positive value implies a codon or amino acid preferred at exonic cores and avoided at the ends. For each species, we then calculate the proportion of codons or amino acids showing significant skew both at 5′- and 3′-ends across all exons and consider this the metric of *cis*-motif usage for that species.

In order to ensure that these trend comparisons are not affected by the different number of exons in different species, for each species, we made a pool of exons and abstracted 5,000 exons from it randomly with replacement (for each repeat of 30 species, 30 data sets were established with each containing 5,000 exons). After 100 repetitions of this sampling process, we obtained the mean usage trends of amino acids and codons for each species by the same method mentioned above. We counted up the number of amino acids or codons that showed a significant rho score in the sample size controlled subsampling and employed this as our metric of the extent of *cis*-motif usage (supplementary table S7.2). We also report results for a sample size uncorrected metric.

Splice-Related Genomic Traits

Based on the data sets of genes saved, we calculated three parameters: *X* (mean CDS length/gene length), *N* (introns per kb exon), and *M* (mean intron size) for each species (supplementary table S7.3).

Phylogenetic Tree with Branch Length

A text file containing an ID list of the 30 species was uploaded to "Taxonomy Browser" of NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi, last accessed January 23, 2014), and then we saved the Taxonomy Common Tree, which has no branch length, in PHYLIP format. To obtain the branch lengths of the phylogenetic tree, a multiple sequence alignment was needed. We searched for candidate orthologs through the orthologous database OrthoDB (http://cegg.unige.ch/orthodb7, last accessed January 23, 2014) and HomoloGene (http://www.ncbi.nlm.nih.gov/homologene, last accessed January 23, 2014) of

NCBI, by taking gene function (related with temperature, air pressure, oxygen concentration, or acid-base properties) into consideration. Ten orthologous genes (supplementary table S7.4), conserved in Eukaryotes, were finally used to make alignments by M-Coffee (http://tcoffee.crg.cat/apps/tcoffee/do:mcoffee, last accessed January 23, 2014) separately. All the ten alignments were merged into one. Gblocks (Castresana 2000; Talavera and Castresana 2007) was employed to eliminate/minimize poorly aligned positions and divergent regions (Supplementary material S1) and converted from Newick format into Nexus within R (parameters used can be found in Supplementary material S2).

Through the RelTime application (Kumar et al. 2012; Tamura et al. 2012, 2013), a phylogenetic tree with branch lengths was constructed by loading the taxonomy common tree and the alignment into MEGA (version 6). We tested the correlation of results from two models (Jones–Taylor–Thornton model and WAG [Whelan and Goldman] model) when selecting "Gamma Distributed" of "Rates and Patterns" and other default parameters. There is a very strong correlation between the branch length estimates from the two models (Spearman correlation: $\rho = 0.9970$, $P = 4.17 \times 10^{-65}$; supplementary fig. S2). We regarded the mean of the results as the final branch lengths (supplementary fig. S3).

Correlation between Amino Acid/Codon Usage Trends and the Genomic Traits after Phylogenetic Correction

The application "Continuous" of BayesTraits (Pagel 1999) was used to study correlations between amino acid/codon usage trends and the genomic traits by Markov chain Monte Carlo method. According to the suggestion from the manual of BayesTraints, we abstracted the last harmonic mean from the result file, and took it as an estimation of marginal likelihood, to calculate the "Log BF" value and further test whether there is strong evidence for the correlation after phylogenetic correction.

Correlation Analysis of $N_e\mu$ Values in Phylogenetic Manner

To calculate $N_e\mu$ values of the species, several R packages (ape [Paradis et al. 2004], PopGenome [Pfeifer et al. 2014], adegenet [Jombart 2008; Jombart and Ahmed 2011], pegas [Paradis 2010], Geiger [Harmon et al. 2008]) and DnaSP (Rozas and Rozas 1995; Librado and Rozas 2009) were used to analyze the published allelic sequences from "PopSet" (http://www.ncbi.nlm.nih.gov/popset, last accessed January 23, 2014) and "Nucleotide" (http://www.ncbi.nlm.nih.gov/nuccore, last accessed January 23, 2014) database of NCBI. Intron sequences were considered as candidates first and, if there is no intron sequence, CDS were chosen for the analysis in which only synonymous sites number, as the segregating sites number, were input the program (supplementary table S8.1). Finally, $N_e\mu$ values ($N_e\mu$ for per site from Pi, $N_e\mu$ for per site from S, and $N_e\mu$ for per site from Eta) of each species were obtained for further correlation analysis (supplementary table S8.2). $N_e\mu$ values from this study were compared, by Spearman's correlation and SMA regression of R package "lmodel2," with the $N_e\mu$ values (Lynch and Conery 2003) published previously (supplementary table S8.3).

By using BayesTraits, in phylogenetic manner, we did correlation analysis of $N_e\mu$ values (both the values from our study and from Lynch and Conery's study) with amino acid/codon usage trends and three intronic dimensions (mean CDS length/gene length, intron density, and mean intron size) (supplementary table S8.4).

Comparison of Selection on Synonymous Mutations with Different Flanking Intron Size

A list of human–macaque orthologs was obtained from ENSEMBL (Flicek et al. 2014). Only those defined as 1:1 orthologs were employed. The respective genes were extracted from human CDS build GRCh37.74 and Macaque build MMUL_1.74. These were aligned using MUSCLE 3.8.31 (Edgar 2004) at the protein level, the nucleotide alignment being built from the protein alignment using a custom script (AA2NUC). Exon and intron sizes for the relevant human genes were obtained through ENSEMBL. Any gene whose CDS length did not match that specified in the BioMart (Kasprzyk 2011) derived annotation file was excluded. The alignment of the exons was derived from the exon dimensions specified (naturally with allowance for indels). Only internal (not first or last exons) exons from the macaque–human comparison were employed.

We considered only exons longer than 2×69 bp and considered the 5′ 69 bp as the 5′-end and 3′ 69 bp as the 3′-end. The alignments were masked with two consensus ESE candidate data sets, INT3 and INT3-400, these being intersect data sets between four high coverage databases of putative ESE sequences (Cáceres and Hurst 2013). One of the data sets presents a large sample of putative ESEs and a second ($N = 400$) top hit sample. As these are nonindependent the intersect data sets employ either the full sample (INT3) or the reduced sample (INT3-400). We could thus, employing these two separately, define sites that were ESE and sites that were possibly not ESE (although as these two sets were conservative, there are likely to be true ESEs in the non-ESE class of sequence). For both ESE and non-ESE masking of the alignments, we then concatenate all exon ends as a function of 20 different flanking intron sizes, thereby making estimation of Ks less noisy. We also compared Ks of exon cores (69 bp of core region in each exon) as a function of the size of neighboring introns after concatenating core sequences in each bin. For each of the sets of concatenated exon ends and cores, both ESE and non-ESE, we estimate Ks using PAML (version: PAML 4.7, Default parameters are used, codon model = 2) (Yang 2007).

To exclude the possibility that any trends seen are not artifacts of skewed nucleotide content between ESE and non-ESE sequence, we generated pseudo-ESE sets containing the same number of random hexamers with, on average, the same nucleotide content as each ESE set. Then, the same test as above was performed 100 times repeatedly for each

pseudo-ESE set. The average value with standard error bar from these nucleotide controls is displayed in the plots (fig. 1*A* and *B*).

Partial Correlation between ESE Density and Three Intronic Dimensions

ESE density and three intronic dimensions (mean intron size, intron density, and intron number) were obtained using custom perl scripts. When two of the intronic dimensions are controlled, partial correlation between ESE density and another intronic dimension was analyzed by R program, pcor.test (Kim and Soojin 2006) (supplementary table S2.1).

Correlation between Flank ESE Constraint and Three Intronic Dimensions

Based on the alignment data set of human–macaque orthologs, Ks core and Ks ESE flank (both 5′ 69 bp and 3′ 69 bp, exons are shorter than 138 bp were all regarded as flank region), of each gene, were calculated after concatenating all flank and core sequences. We set up three criteria for concatenating sequence to select genes for correlation analysis (1. ESE flank $>$ 102 bp, Core region $>$ 102 bp; 2. ESE flank $>$ 150 bp, Core region $>$ 150 bp; 3. ESE flank $>$ 201 bp, Core region $>$ 201 bp). Then, three intronic dimensions (mean intron size, intron density, and intron number) and Flank ESE constraint of each gene were obtained by our perl script. For INT3 and INT3_400 data sets, we explored the correlation between Flank ESE constraint (defined in results) and three intronic dimensions, considering 5′- and 3′-ends of exons separately, by partial Spearman's correlation (supplementary table S3.2).

We evaluate the net effect of flanking intron size (constraint and increased density) by calculating the proportion of synonymous sites under ESE-related constraint at exon flanks as flank ESE constraint × ESE density. Instead of using the conservative binning method ($N = 20$), we calculated the regression line of logarithm value of flank intron size versus ESE constraint and for each exon individually calculate ESE density × constraint, where constraint is estimated through interpolation of this regression line, given the intron size. Then, we examine whether the Spearman's rank correlation between ESE density × constraint and the logarithm value of intron size is significant.

Furthermore, Goodman and Kruskal's gamma, by using program "rcorr.cens" from R package "Hmisc" (http://biostat.mc.vanderbilt.edu/Hmisc, https://github.com/harrelfe/Hmisc, last accessed January 23, 2014) was carried out in above analysis to avoid affects of tied observations. *P* value, which shows whether Goodman and Kruskal's gamma is significant, comes from $p = (n + 1)/(m + 1)$ where n is the number of gamma values calculated after randomly shuffled the variables representing flank ESE constraint and meanwhile greater than the observed gamma and m is 1,000, this being the number of times of shuffling (supplementary table S3.1).

To make sure the above result is not affected by sample size artifacts, we did a resampling test by abstracting genes by the number in the 150-bp cutoff group from the gene pool in the 120-bp cutoff group and repeated the two types of correlation analysis 1,000 times. We report the proportion of random subsamplings that still provide a significant correlation prior to multitest correction.

Comparison of ESE Density between Second Exons and Last but One Exons

In the human gene data set, we selected from genes with four or more exons, second exons and last but one exons, which are all greater than 138 bp. We calculated 5′-exon end ESE density and 5′-flank intron size of these two exon categories in each gene. To control for the effect of flank intron size, we analyzed the residuals from loess regression of 5′-end ESE density predicted by 5′-intron size (supplementary table S6.1). Both analyses were repeated for the two ESE data sets (INT3 and INT3_400). Significance was assayed through a binomial test counting the absolute number of genes having a higher density at the second exon than the last but one, versus the opposite. If ESE density was no different, these were ignored.

Relationship between Transcriptional Elongation Rate and ESE Density

We used publicly available data from a genome-wide elongation rate study (Veloso et al. 2014) to investigate the relationship of ESE density with transcriptional elongation rate (around 450 genes were selected due to requirement of ESE density calculation; supplementary table S4.1) and also correlated the elongation rates with several genic dimensions used in our study (supplementary table S4.2).

Supplementary Material

Supplementary tables S1–S8, figures S1–S3, and files S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a University Research Studentship from the University of Bath to X.M.W. and a Wolfson Royal Society Research Merit Award to L.D.H. The work was funded in part by Medical Research Grant MR/L007215/1. The authors thank Mike Lynch for access to data and advice regarding data.

References

Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z. 2010. A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J Biol Chem.* 285:28741–28748.

Bell MV, Cowper AE, Lefranc MP, Bell JI, Screaton GR. 1998. Influence of intron length on alternative splicing of CD44. *Mol Cell Biol.* 18:5930–5941.

Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem.* 270:2411–2414.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* 25:106–110.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hébuterne X, et al. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet.* 43:242–245.
Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* 14:R143.
Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol.* 62:89–98.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.
Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.
Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Mol Biol Evol.* 31:1402–1413.
Daubin V, Moran NA. 2004. Comment on "The origins of genome complexity.". *Science* 306:978; author reply 978.
Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev.* 12:640–649.
Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol.* 40:308–317.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol.* 27:177–192.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007–1013.
Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.
Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A.* 102:16176–16181.
Frank SA. 2007. Maladaptation and the paradox of robustness in evolution. *PLoS One* 2:e1021.
Gartner JJ, Parker SC, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, et al. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A.* 110:13481–13486.
Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55:2161–2169.
Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* 6:1197–1211.
Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. *Bioinformatics* 24:129–131.
Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J Mol Evol.* 63:174–182.
Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol.* 158:573–597.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2:13–34.
Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405.
Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071.
Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol.* 53:290–298.
Kandul NP, Noor MA. 2009. Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC Genet.* 10:67.
Kasprzyk A. 2011. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* 2011:bar049.
Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360–1374.
Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3:e42.
Kim S-H, Soojin VY. 2006. Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. *Mol Biol Evol.* 23:1068–1075.
Klinz FJ, Gallwitz D. 1985. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 13:3791–3804.
Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 28: 2685–2686.
Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 9: e1003527.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A.* 108:11093–11098.
Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A.* 98: 11193–11198.
Luehrsen KR, Walbot V. 1992. Insertion of non-intron sequence into maize introns interferes with splicing. *Nucleic Acids Res.* 20: 5181–5187.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates; Basingstoke: Palgrave [distributor].
Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.
Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314:1930–1933.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst.* 23:263–286.
Ohta T. 1996. The current significance and standing of neutral and neutral theories. *Bioessays* 18:673–677; discussion 683.
Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.
Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26:419–420.
Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol.* 24:1600–1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.
Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol.* 31:1929–1936.
Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6:e1001236.
Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet.* 24:590–594.
Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515:261–263.
Rozas J, Rozas R. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Comput Appl Biosci.* 11:621–625.
Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond B Biol Sci.* 349:241–247.
Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J. 2006. An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A.* 103:5006–5011.
Spingola M, Grate L, Haussler D, Ares M Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* 5:221–234.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.
Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci U S A.* 109:19333–19338.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol.* 30:2725–2729.
Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, Ljungman M. 2014. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* 24:896–905.
Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9:R29.
Whitney KD, Garland T Jr. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet.* 6:e1001080.
Wu X, Tronholm A, Cáceres EF, Tovar-Corona JM, Chen L, Urrutia AO, Hurst LD. 2013. Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga *Ectocarpus* are common and resemble those seen in humans. *Genome Biol Evol.* 5:1731–1745.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Chapter IV.

Determinants of the usage of splice-associated *cis*-motifs predict the distribution of human pathogenic SNPs

Published manuscript:

XianMing Wu, Hurst LD. Determinants of the usage of splice-associated *cis*-motifs predict the distribution of human pathogenic SNPs. ***Molecular Biology and Evolution*.** 2015 Nov 8.

Contributions

All analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst.

MBE Advance Access published December 8, 2015

Determinants of the Usage of Splice-Associated *cis*-Motifs Predict the Distribution of Human Pathogenic SNPs

XianMing Wu[1], Laurence D. Hurst*[,1]

[1]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, Somerset, United Kingdom

***Corresponding author:** E-mail: l.d.hurst@bath.ac.uk.

Associate editor: Jianzhi Zhang

Abstract

Where in genes do pathogenic mutations tend to occur and does this provide clues as to the possible underlying mechanisms by which single nucleotide polymorphisms (SNPs) cause disease? As splice-disrupting mutations tend to occur predominantly at exon ends, known also to be hot spots of *cis*-exonic splice control elements, we examine the relationship between the relative density of such exonic *cis*-motifs and pathogenic SNPs. In particular, we focus on the intragene distribution of exonic splicing enhancers (ESE) and the covariance between them and disease-associated SNPs. In addition to showing that disease-causing genes tend to be genes with a high intron density, consistent with missplicing, five factors established as trends in ESE usage, are considered: relative position in exons, relative position in genes, flanking intron size, splice sites usage, and phase. We find that more than 76% of pathogenic SNPs are within 3–69 bp of exon ends where ESEs generally reside, this being 13% more than expected. Overall from enrichment of pathogenic SNPs at exon ends, we estimate that approximately 20–45% of SNPs affect splicing. Importantly, we find that within genes pathogenic SNPs tend to occur in splicing-relevant regions with low ESE density: they are found to occur preferentially in the terminal half of genes, in exons flanked by short introns and at the ends of phase (0,0) exons with 3′ non-"AGgt" splice site. We suggest the concept of the "fragile" exon, one home to pathogenic SNPs owing to its vulnerability to splice disruption owing to low ESE density.

Key words: **pathogenic SNPs, splicing *cis*-motif, splice site, exonic splicing enhancer.**

Introduction

Although it is clear that many disease-causing mutations are nonsense or nonsynonymous changes within exons, that many synonymous mutations also cause disease (Chamary et al. 2006; Sauna and Kimchi-Sarfaty 2011; Hunt et al. 2014; Bali and Bebok 2015) suggests that mechanisms beyond replacement of one amino acid for another (or for a stop), can be important. That synonymous mutations cause disease also suggests that many nonsynonymous mutations might have their effects for reasons other a slightly altered amino acid content of proteins.

One of the predominant mechanisms by which synonymous mutations cause disease (and affect fitness more generally) is via modulation of splicing (Faustino and Cooper 2003; Chamary et al. 2006). Indeed, it is estimated that of known disease-associated synonymous variants perhaps over 90% impact splicing (Mort et al. 2014). In some cases, the effect is a simple disruption of the splice site. It is for example estimated that 15% of splice site mutations could lead to human genetic disease (Krawczak et al. 1992). However, splice sites, while important, do not contain all the information for splicing in humans (Wang and Burge 2008). Indeed, in the human genome approximately 50% of the information defining splice sites is in *cis*-motifs, typically in close proximity to the splice sites (Lim and Burge 2001). More generally, the behavior of these *cis*-motifs is dependent on the intragene location, with one motif having different activity dependent on local context and position within the exon (Wang and Burge 2008; Ke et al. 2011). Detailed studies suggest that for some exons, 30% of individual mutations in a given exon can affect the splice pattern (Pagani et al. 2005). In turn, it is expected that disruption of *cis*-motifs might also cause disease, a prediction borne out by the evidence (Nissim-Rafinia and Kerem 2002; Faustino and Cooper 2003). For instance, a splicing enhancer disruption in exon 51 of *FBN1* gene relates to Marfan syndrome (Caputi et al. 2002) and a point mutation in exon 7 of *SMN2* gene is associated with Motor Neurone Disease (Gavrilov et al. 1998; Cartegni and Krainer 2002; Wirth et al. 2006).

The possible selective importance of *cis*-motifs is underscored by the observed selection to preserve them at the terminal regions of exons. Many studies have reported that selective constraints are more common at the ends of exons due to the splicing control, such that new mutations at exon ends are likely to be eliminated by purifying selection, even in species with low effective population size (Majewski and Ott 2002; Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006, 2007; Cáceres and Hurst 2013; Wu and Hurst 2015). To date, two important exonic splicing control elements, serving as enhancers (exonic splicing enhancers, ESEs) (Blencowe 2000) and silencers (exonic splicing silencers, ESSs) (Amendt et al. 1995; Kan and Green 1999), have been investigated. These *cis*-motifs generally reside within exon ends and function by interacting with certain regulators (SR proteins and hnRNP) (Zheng et al. 2000; Rowen et al. 2002). With little evidence that ESS motifs are under purifying

Article

selection (Chamary et al. 2006; Parmley and Hurst 2007; Parmley et al. 2007), while ESEs generally are, we here concentrate our attention on ESEs. It is estimated that about 4–5% of synonymous mutations are under selection in humans because they disrupt ESEs (Cáceres and Hurst 2013). In particular, we ask whether the biases in intragene distribution of ESEs predicts in any manner biases in the intragene distribution of pathogenic single nucleotide polymorphisms (SNPs). Such a covariance would lend support to the postulate that splice disruption is an important mechanism for pathogenesis (Cartegni et al. 2002). It would also suggest that we ignore synonymous SNPs for diagnostics at our peril.

The notion that ESE location might predict disease causing SNP distributions has some support. Recently, splice-affecting SNPs (likely to be enriched for pathogenic mutations) were found to be significantly enriched at exon ends (Woolfe et al. 2010), the hot spots of ESEs (Nelson and Green 1988; Lavigueur et al. 1993; Graveley et al. 1998; Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006, 2007; Cáceres and Hurst 2013). Therefore, for understanding patterns of molecular evolution, for understanding splice control and for understanding how and why mutations cause disease, it is helpful to have a robust understanding of the relationship between ESE usage and distribution of disease-causing mutations. Here, we address this by observing the coupling of ESE density with relative abundance of exonic pathogenic SNPs.

ESE usage within and between genes is known to vary with numerous factors. As noted above, ESEs tend to be at their most dense in the terminal (up to 70–100 bp) portion of exons (Fairbrother et al. 2004; Cáceres and Hurst 2013). Moreover, ESE usage has recently been shown to be higher in the 5′-exons compared with 3′-exons (Wu and Hurst 2015) and in exons flanking large introns (Dewey et al. 2006; Cáceres and Hurst 2013; Wu and Hurst 2015), as well as in genes with high exon density (number of exons per coding bp) (Wu and Hurst 2015). These covariates were conjectured to reflect selection to minimize the impact of possible decoy splice sites (Wu and Hurst 2015). This selection, it was conjectured, would be more common for earlier exons (more downstream splice sites), exons flanked by larger introns (more possibilities for decoys) and in genes with more exons (more splice sites, hence more decoys for any given splice site; Wu and Hurst 2015).

ESE usage is also a function of the flanking splice site (Berget 1995; Graveley 2000), a higher ESE density being considered necessary to support the reinforcement of the flanking "weak" splice sites (Fairbrother et al. 2002; Dewey et al. 2006; Plass et al. 2008; Cáceres and Hurst 2013). This too can fit into the broader decoy splice site model, in the sense that if the weak splice site is not found a decoy (inappropriate) one might be. Recently, we found usage of splicing *cis*-motifs correlates well and positively with the usage of tetra-nucleotide splice sites "AGgt" and "agGT" (nucleotides of exons in upper case, nucleotides of introns in lower case) across 30 species (Wu and Hurst 2015).

In further examination we noted, serendipitously, that exon phase appears to be a predictor of ESE density (for evidence see below). Phase here refers to where in the codon the splice site hits, a phase zero exon end being one cut between whole codons. The three possible phases are not found in equal proportions (Fedorov et al. 1992; Long et al. 1995; Ruvinsky et al. 2005). Furthermore, in phylogenetically manner, we observe a significant and negative correlation between proportion of phase zero splice site and usage of ESEs (table 1). To date, we have no good explanation as to why phase might matter, but for the purposes of this article such considerations are not relevant. We simply consider it to be a correlate to ESE density.

Given the above we hence consider the intragenic location of pathogenic SNPs and five possible covariates: 1) relative position in exons, 2) relative position in genes, 3) flanking intron size, 4) usage of splice sites, and 5) exon end phases. For each we consider whether pathogenic SNPs are more or less likely to occur where ESEs are more or less common. As many genes are not associated with diseases (either because they are too essential or too unimportant), we consider the trends within the class of genes within which disease-associated SNPs are found. We start, however, by asking whether disease causing genes are at all unusual in the exon–intron architecture.

Results

Disease-Associated Genes Tend to Have More Exons

Before considering intragenic trends we first ask whether genes that contain pathogenic SNPs are intrinsically those most likely to be affected by splice-related mechanisms. Were pathogenic SNPs to have their effect via missplicing, we might expect that genes with more exons (more splice sites) tend to be disease-associated. According to disease-related information of sequence variation in Clinvar database (http://www.ncbi.nlm.nih.gov/clinvar/, last accessed November 24, 2014; Landrum et al. 2014), we established a data set of 9,818 pathogenic SNPs coming from missense and silent mutations (not nonsense) (supplementary table S1, Supplementary Material online). Of these SNPs 8,250 (84%)

Table 1. *Cis*-Motif Usage Correlates Significantly with Proportion of Phase Zero Splice Sites.

	All Exons (AA)	All Exons (codon)	Random 5,000 Exons (AA)	Random 5,000 exons (codon)
Log BF[a] (Y∼$P_{phase\text{-}zero}$)	140.6781	103.9811	82.6049	114.0758
R Trait 1 2[b]	> 0	>0	>0	>0

NOTE.—Y, proportion of amino acids/codons showing significant trends; $P_{phase\text{-}zero}$, proportion of intercodon splice site.

[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)]−log [harmonic mean (simple model)]), All Log BF values in the table are > 10, so the evidences of all correlations are very strong.

[b]BayesTraits parameter "R Trait 1 2" can indicate whether the correlation is positive (> 0) or negative (<0).

exist in internal exons. From UCSC (Karolchik et al. 2004) (http://genome.ucsc.edu/cgi-bin/hgTables, last accessed November 24, 2014), 1,747 gene sequences that contain these pathogenic SNPs were derived. Note that here and elsewhere we exclude nonsense SNPs, as their likely mode of pathology is probably not splice related (but see below).

Consistent with the splice decoy model for determining the genic richness of splicing-related *cis*-motifs, we found that exon number correlates significantly with ESE usage (Wu and Hurst 2015). In turn, we find that the absolute number of exons in disease genes is higher than that in nondisease gene (Mann–Whitney *U* test: $P = 5.87 \times 10^{-119}$; median of number of exons in disease genes = 12, median of number of exons in nondisease genes = 8). However, under the null that all bases in all coding sequences (CDSs) have the same likelihood of causing disease, we expect long genes to be more likely to be disease associated (just because they have more base pairs). This is indeed the case (Mann–Whitney *U* test: $P = 2.41 \times 10^{-103}$; median of CDS size in disease genes = 1,689 bp, median of CDS size in nondisease genes = 1,248 bp). So here we ask whether disease-associated genes have more exons when CDS length is controlled.

As the relationship between CDS length and number of exons is not linear, we employ a Mann–Whitney *U* test to analyze residuals of a loess regression for number of exons in all genes (disease and nondisease genes) against CDS length. We find that the number of exons in disease-associated genes, controlling for CDS size by this method, is higher than that of nondisease genes (Mann–Whitney *U* test: $P = 8.11 \times 10^{-23}$, median of residuals for disease genes: 0.921, median of residuals for nondisease genes: 0.146) (supplementary table S2, Supplementary Material online). One interpretation of this is that the more exons in a gene the more likely an inappropriate splice event might take place. However, as exon size and expression level covary, this interpretation is by no means unique.

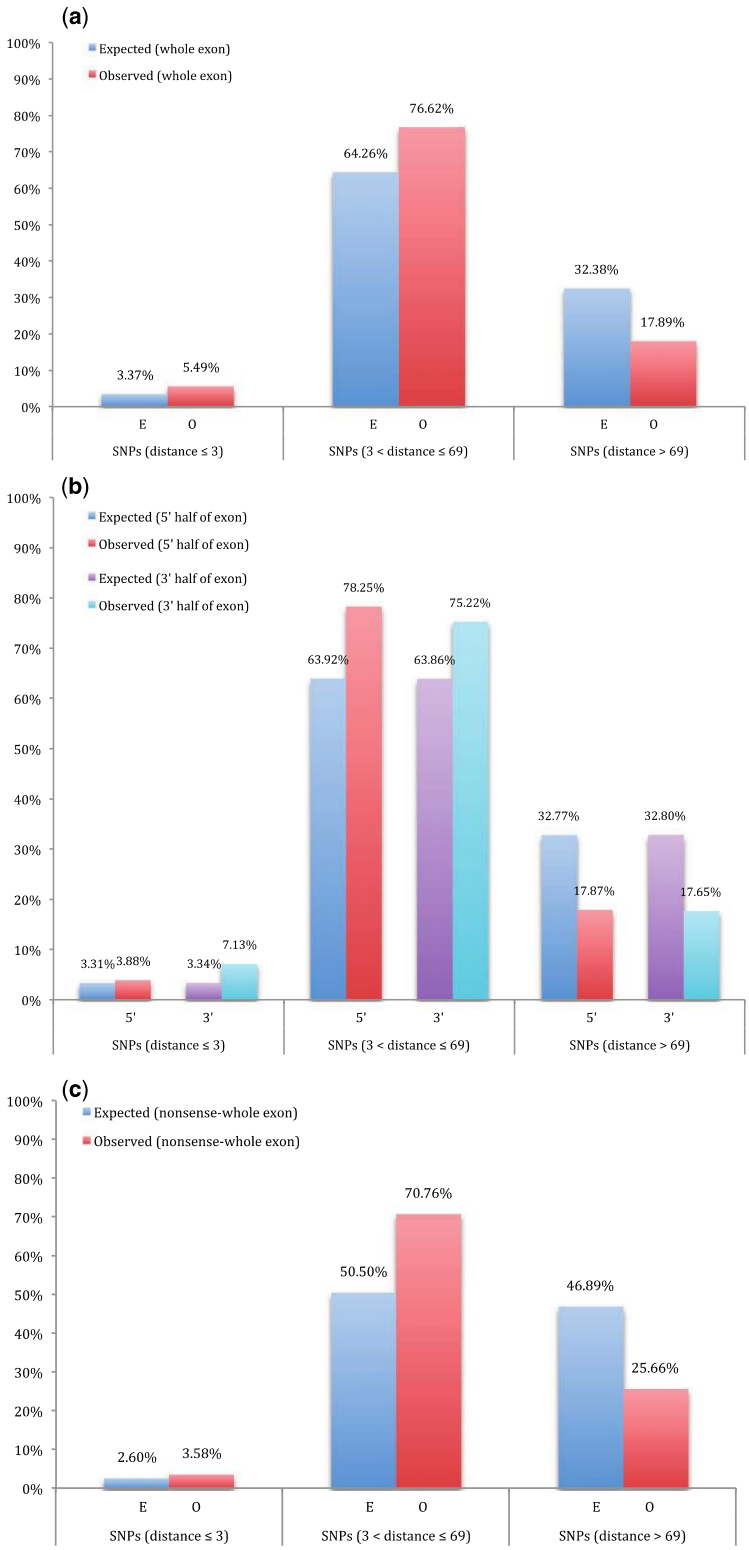
Both Splicing *cis*-Motifs and Pathogenic SNPs Tend to Be Present at Ends of Exons

We now turn to intragenic predictors of where non-nonsense pathogenic SNPs are over or underrepresented. It has been found that *cis*-exonic splice control elements, such as ESEs, tend to enriched at the ends of exons (Nelson and Green 1988; Lavigueur et al. 1993; Graveley et al. 1998; Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006, 2007; Cáceres and Hurst 2013). This is at least part of the explanation for reduced substitution rates and rarity of SNPs at the ends of exons (Majewski and Ott 2002; Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006, 2007; Cáceres and Hurst 2013; Wu and Hurst 2015). Consistent with this, splice-affecting variants (many of which we expect to be pathogenic), are significantly enriched at both ends of exons (Woolfe et al. 2010). We thus ask whether pathogenic SNPs in internal exons (not first and last exons) are more common at exon ends.

Of the 8,250 pathogenic SNPs in internal exons, the great majority (76.62%) are within 3–69 bp from the ends of the internal exons. This statistic, however, is uninformative without assessment of relative enrichment. To this end, we consider enrichment of SNPs in three domains: within 3 bp of splice sites; between 3 and 69 nt from exon ends; all other sequence from internal exons, that is, exon core. Comparing the distribution of pathogenic SNPs in the exons against the expected distribution in the same SNP bearing exons, we find that splice sites ($\leq$3 bp) are greatly enriched for pathogenic SNPs (Observed: 5.49%, Expected: 3.37%), as are exon terminal domains (3–69 bp, Observed: 76.62%, Expected: 64.26%). Given this enrichment, it is inevitable that exon cores are relatively underrepresented (Observed: 17.89%, Expected: 32.38%; fig. 1*a*). This deviation is highly significant ($\chi^2 = 841.64$, df = 2, $P < 1.74 \times 10^{-183}$) (supplementary table S3, Supplementary Material online).

To examine whether this biased distribution affects both 5′- and 3′-halves of exons, we perform a similar enrichment analysis considering 5′- and 3′-exon ends separately. The same pattern of significant deviation in pathogenic SNP distribution can still be observed: in both 5′ ($\chi^2 = 410.62$, df = 2, $P < 6.85 \times 10^{-90}$) and 3′ ($\chi^2 = 551.74$, df = 2, $P < 1.55 \times 10^{-120}$) half of internal exons, splice sites ($\leq$3 bp) and exon terminal domains (3–69 bp) are preferred regions; and exon cores ($>$ 69 bp) are relatively avoided (fig. 1*b*) (supplementary table S3, Supplementary Material online).

We can go further and attempt to estimate the proportion of non-nonsense pathogenic SNPs that have their effects via splicing. If we assume that mutations in exon cores never disrupt splicing, then pathogenic SNP rates in cores provide a background splice-unconnected rate. This assumption is likely to be wrong, not least because splice disrupting mutations in exon cores are described (Woolfe et al. 2010), but renders the method conservative. First, in all disease genes, we estimate the background nonsplicing-related pathogenic SNP rate by asking about the frequency of pathogenic SNPs in exon core regions, defined as the central 100 bp of internal exons longer than 300 bp (there are 707 internal exons longer than 300 bp). The 70,700 (*L*) bp of 100 bp core regions contain 124 (*N*) pathogenic SNPs, which gives us $N/L = 0.001754$ SNPs per bp. If 69 bp is considered as exon "end" cutoff, we can then calculate how many pathogenic SNPs we would expect in exon flanks, assuming flanks behave like these cores, this being the core rate per base pair multiplied by the number of exon flank base pairs (2,734,972 bp). This yields an estimate of 4,797 pathogenic SNPs. We actually observe a total of 6,774 SNPs in these exon flank domains (N.B. there are 1,476 pathogenic SNPs in regions outside of these flanks, so 8,250 (6,774 + 1,476) pathogenic SNPs in internal exons). If we assume that the excess at exon ends (6,774 — 4,797 = 1,977 bp) affect splicing, we can calculate the proportion, at 69 bp cutoff exon ends, of all pathogenic SNPs that affect splicing as 1,977/8,250 = 23.97% (supplementary table S4, Supplementary Material online). This method is quite robust to definition of exon end. If we consider different definitions of splice affected exons exon ends in a range from 50 to 100 bp, estimates range from 20% at 50 bp definition of exon end to 27% at 100 bp (supplementary table S4, Supplementary Material online).



The above method is conservative in assuming that splice disrupting mutations only ever occur at exon ends. However, prior estimates (Woolfe et al. 2010) suggest that about 25% of mutations in exon core regions also affect splicing. A more liberal estimate then supposes that the overall background (nonsplice dependent rate) rate is $0.75 \times N/L = 0.001315$. This being so, with a total number of base pairs in internal exons of all pathogenic genes ($T = 3\,527\,520$ bp), we expect $T \times 0.001315$ (=4,639) pathogenic SNPs to not exercise their effects via splicing. With a total of 8,250 pathogenic SNPs in all internal exons, the proportion that do exercise their effects via splicing defects we estimate to be (8,250 — 4,639)/8,250 = 43.77%. Both the conservative (~20%) and liberal estimates (~43%) support the hypothesis that splice disruption is a major cause of pathology.

Note that we have assumed that nonsense mutations have their effects via the introduction of premature stop codons. However, in principle a stop could also disrupt ESEs and lead to disrupted splicing, possibly leading to the noninclusion of the exon with the mutation. We can then ask about the proportion of all SNPs (including nonsense mutations; $N = 10{,}764$ SNPs in internal exons; supplementary table S3, Supplementary Material online). Repeating the above calculations for all SNPs distribution in internal exons, we find similar and significant deviated patterns (supplementary fig. S1, Supplementary Material online). These numbers are highly similar to those obtained using just the non-nonsense SNPs. Although nonsense mutations associated with pathology are relatively rare (2,514 SNPs) (supplementary table S3, Supplementary Material online), these too show an excess at exon ends (fig. 1*c*).

Pathogenic SNPs Are more likely to Occur in the Terminal Half of Genes

It was previously observed that exons in 5′-positions in genes have higher ESE density than more terminal exons. Comparing second exons to last but one exons within the same gene, for example, it was observed that there is a 2-fold greater ESE density in the former (Wu and Hurst 2015). This it was suggested might reflect the fact that early exons, by definition, have more downstream splice sites than do later ones and hence have a larger number of potential decoy splice sites. Does then 5′- to 3′-position predict pathogenic SNP density? Expectations here are unclear. It might be that 5′-pathogenic SNPs might be more common as they might have a more drastic effect on the subsequent protein. Alternatively, with high ESE density, a given SNP might be less prone to disrupting splicing, the ESEs providing some degree of resilience. A further confounding factor is that if a SNP has a major effect it might be an embryonic lethal and hence not be classified as pathogenic.

In the first instance we divide the genes (not the CDS), into the absolute 5′-first half and the 3′-half (from ATG to stop). We observe that pathogenic SNPs are more likely to occur in the rear (3′) half section of genes (5′-first half: 3,393, 3′-rear half: 6,425; ratio = 6,425/3,393 = 1.89; Binomial Distribution P value: 1.74×10^{-209}) (supplementary table S1, Supplementary Material online). However, this analysis neglects the influence of any differential distribution of CDS sequence between first half and terminal half of genes. To address this, we performed a simulation in which we randomly select a pseudo pathogenic SNP (the same nucleotide as the mutating one) in each disease-causing gene. Then, for the whole disease-causing gene data set, we calculate the ratio of number of pseudo SNPs in the 3′-half to that in the 5′-half and take this as the expected ratio to compare with the real observation. After 100 repetitions of the process, we found all simulated ratios (number of 3′-half pseudo SNPs/number of 5′-half pseudo SNPs, mean = 1.7) are less than the observed ratio (ratio = 1.89) ($P = 0.0099$ by 100 randomizations; supplementary table S5, Supplementary Material online).

One might suggest that this need not reflect splicing defects, but rather the lesser impact of mutations in the 3′-half of the CDS. This, however, appears not to be the case. If we consider the distribution in CDS sequences, then there is no significant difference after Bonferroni correction (supplementary table S1, Supplementary Material online). This indicates that relative position within the CDS is not so important, while relative position in the unspliced RNA is. This would be consistent with splice defects being of relevance, but such an interpretation is by no means unique.

Pathogenic SNPs Preferentially Reside in Proximity to Shorter Flanking Introns

In the human genome, ESE density tends to be higher in the exons flanked by larger introns (Dewey et al. 2006; Cáceres and Hurst 2013; Wu and Hurst 2015). Here, to test if flanking intron size affects distribution of pathogenic SNPs, we consider the relationship between the density of pathogenic SNPs (Dpi = number of pathogenic SNPs/exon length) and flanking intron size.

First, as many exons have no pathogenic SNP, we perform this test only for the exons with pathogenic SNPs. We observe a weak but significant negative correlation between Dpi (supplementary table S6, Supplementary Material online) and the log of flanking intron size (here, "flanking intron" means the "nearest intron") (Spearman correlation, $\rho = -0.085$,

Fig. 1. Pathogenic SNPs are enriched close to exon junctions. (*a*) Of 8,250 pathogenic SNPs in internal exons, the great majority (76.62%) are within 3–69 bp from the exon ends. We consider enrichment of SNPs in three domains: 1) splice sites (≤3 bp) are greatly enriched for pathogenic SNPs (Observed: 5.49%, Expected: 3.37%); 2) Pathogenic SNPs have significant preference at exon terminal domains (3–69 bp, Observed: 76.62%, Expected: 64.26%). 3) Distribution of pathogenic SNPs in exon cores are relatively underrepresented (Observed: 17.89%, Expected: 32.38%). ($\chi^2 = 841.64$, df = 2, $P < 1.74 \times 10^{-183}$). (*b*) The same pattern of significant deviation in pathogenic SNPs distribution are observed for: 1) 5′-half of internal exons ($\chi^2 = 410.62$, df = 2, $P < 6.85 \times 10^{-90}$); 2) 3′-half of internal exons ($\chi^2 = 551.74$, df = 2, $P < 1.55 \times 10^{-120}$). (*c*) Distribution of nonsense pathogenic SNPs in internal exons are similar to that of non-nonsense mutations ($\chi^2 = 455.37$, df = 2, $P < 1.31 \times 10^{-99}$).

Table 2. Evidence for Correlation between Proportion of Phase Zero Splice Sites and Splice-Related Genomic Traits.

	X	*N*	*M*
Log BF [a] ($P_{phase\text{-}zero}$ ~ Splice-related genomic traits)	**58.5625**	**45.2272**	**26.6799**
R Trait 1 2[b]	**> 0**	**<0**	**<0**

NOTE.—X, mean CDS length/gene length; N, introns per kb exon; M, mean intron size; $P_{phase\text{-}zero}$, Proportion of phase zero splice site.

[a]Log BF (log Bayes factor) = 2*(log [harmonic mean (complex model)]-log [harmonic mean (simple model)]), All Log BF values in the table are > 10, so the evidences of all correlations are very strong.

[b]BayesTraits parameter "R Trait 1 2" can indicate whether the correlation is positive (> 0) or negative (<0).

$P = 1.14 \times 10^{-8}$). However, this could be partly due to covariance between exon size and flanking intron size. Indeed, our Spearman's correlation shows that exon size correlates well and positively with flanking intron size, not only for pathogenic SNP-containing exons ($\rho = 0.092$, $P = 6.84 \times 10^{-10}$), but also for all exons within disease-causing genes (if there is no pathogenic SNP in the exon: Dpi = 0) ($\rho = 0.037$, $P = 7.94 \times 10^{-10}$). Furthermore, as intron sizes are known to covary with gene expression, and gene expression is likely to predict whether a gene is associated with disease (Emilsson et al. 2008), it is also necessary to control for expression level. One way to do this is to perform a covariate controlled analysis, but this depends on the accuracy of the expression data. An alternative method is to perform an intragene analysis, comparing exons within the same gene, which we presume to have the same expression level.

To this end, we set up a simulation in which we randomly select a nucleotide the same as the mutating nucleotide, within the same gene, in internal exons (not including first and last exons) of all disease genes. Then we can ask how many flanking introns (in the same genes as the SNP) of the selected pseudo SNPs are larger than (or equal to) those of real SNPs (binomial test for significance). In every trial of 100 randomization processes, the number of expected flanking introns (flanking intron of pseudo SNPs) that are larger than observed ones (flanking intron of real SNPs) is significantly greater than the number of expected flanking introns that are shorter than observed ones (supplementary table S7, Supplementary Material online, $P = 0.0099$). So, despite covariance with exon size, pathogenic SNPs tend to occur in exons flanked by smaller introns, within the same gene.

To further consider the issue, we considered the density of pathogenic SNPs (Dpi) for first half and terminal half of exons separately (see Materials and Method). We set up a data set of 238 genes that have at least five internal pathologic SNP containing exons. We then perform a Spearman's correlation analysis between Dpi and flanking (nearest) intron size for each gene. We expect that most intragene comparisons will have a negative correlation (as seen in the between-gene comparison). We find that a significant majority of intragene correlations are indeed negative (Negative correlation: 136, positive correlation: 99; Binomial test $P = 2.83 \times 10^{-3}$; supplementary table S8, Supplementary Material online). We conclude that exonic-disease causing SNPs tend to occur in the vicinity of shorter introns and that this cannot be fully explained by certain possible covariant factors, such as exon size, expression level, and flanking intron selection.

Pathogenic SNP Distribution Indicates Enrichment Near Zero Phase and 3′ Non-"AGgt" Splice Sites

Usage of tetranucleotide splice sites AGgt and agGT has been found to correlate well and positively with usage of splicing *cis*-motifs across species (Wu and Hurst 2015). Within each of these splice sites, the 2 nt in upper case come from exons. By serendipity we also find that coding phase has a relationship with the usage of tetranucleotide splice sites. This we found through investigating whether there might be a relationship between phase and specific splice site (AGgt and agGT) usage in the human genome. We discovered there to be a significant relation (for AGgt: χ^2 test: $\chi^2 = 308.18$, df = 2, $P = 1.20 \times 10^{-67}$, For agGT: χ^2 test: $\chi^2 = 139$, df = 2, $P = 6.57 \times 10^{-31}$; supplementary table S9, Supplementary Material online). Through calculating the "fo/fe" value (the ratio of observed number to expected number), we find that AGgt and agGT splice sites tend to be in the phase zero exon ends (supplementary table S9, Supplementary Material online).

We then asked whether splice site phase might also be a predictor of *cis*-motif usage as well as the tetranucleotide splice sites. To test for such a coupling, we performed a phylogenetic correlation analysis employing BayesTraits (Pagel 1999). Across 30 species, we find that exonic splice *cis*-motif usage correlates significantly and negatively (BayesTraits parameter "R Trait 1 2" < 0) with the proportion of phase zero exon ends (table 1). Furthermore, there is also strong evidence for a correlation between proportion of phase zero splice sites and three intronic parameters: *X* (mean CDS length/gene length), *N* (introns per kb exon) and *M* (mean intron size) for each species (table 2; supplementary table S10, Supplementary Material online). The above suggests that there exist trends across species as regards the proportion of a given phase being used and *cis*-motif usage.

Since usage of certain tetranucleotide splice sites (AGgt and agGT) and the proportion of zero phase splice site correlates well with ESE usage, do these two predictors of *cis*-motifs usage predict the occurrence of pathogenic SNPs? We address the question using a randomization approach. There are 1,900 SNPs in exons with agGT splice site and 4,534 SNPs flanked by AGgt. To consider the phase of exon ends, there are 1,958 SNPs in exons that are phase 5′ = 0 and phase 3′ = 0 (i.e., symmetric (0,0) exons). We establish the significance of these figures (observed value) by randomly selecting a nucleotide identical to the mutating nucleotide in each specific gene (this can exclude the influence of gene expression level and biased nucleotide content) and, in turn, calculate its exonic end phases and identify the nucleotide content of flanking tetranucleotide splice sites. For all collected pseudo

pathogenic SNPs, we obtain the numbers (Expected value) of exons with agGT, AGgt, and phase (5′ = 0, 3′ = 0) separately. We perform this 100 times with P value given by $p = (n + 1)/(m + 1)$, where n is the number of expected values calculated after randomly selecting pseudo SNPs and meanwhile greater (for phase)/smaller (for AGgt) than (or equal to) the observed value and m is 100, this being the number of randomization cycles. We observe that pathogenic SNPs are more likely to be in symmetric (0,0) exons ($P = 0.0099$). The usage of the two specific splice sites seems to have different preferences: for AGgt, observed usage is significantly less than that expected ($P = 0.0099$), meanwhile, agGT has no significant usage pattern (supplementary table S11, Supplementary Material online).

We can also attempt to estimate the degree of enrichment. For all disease-causing genes, we add up the length of all 3′-AGgt splice site internal exons (1,962,796 bp, 55.64%), and calculate the total length of other phase internal exons (1,564,724 bp). The number of SNPs in 3′-AGgt exons is 4,330 (52.48%) and that for other splice site exons is 3,920. This suggests significant avoidance of pathogenic SNPs in 3′-AGgt splice site exons (Observed ratio (52.48%)/ Expected ratio (55.64%) = 0.94) ($\chi^2 = 33.33$, df = 1, $P < 7.80 \times 10^{-9}$) (supplementary table S12, Supplementary Material online). Similarly, we did enrichment analysis for pathogenic SNPs in symmetric (0,0) internal exons. The length of all such (0,0) exons is 805,767 bp (22.84%) and the total length of other phase internal exons is 2,721,753 bp. As above, the number of SNPs in phase (0,0) exons is 1,958 (23.73%) and that in other phase exons is 6,292. This suggests a weak but significant enrichment of pathogenic SNPs in phase zero exons (Observed ratio (23.73%)/ Expected ratio (22.84%) = 1.04) ($\chi^2 = 3.72$, df = 1, $P < 0.05$) (supplementary table S12, Supplementary Material online).

Therefore, exons with 3′-splice site being not AGgt and with zero phase at both ends are more likely to contain pathogenic SNPs. As both phase zero exon ends and 3′ non-AGgt splice sites are characteristics of exons with low ESE density, pathogenic SNPs and ESE usage appear to be anti-correlated.

ESE Density and Pathogenic SNP Density Negatively Covary

Above we have considered several predictors of ESE densities to ask whether in any manner they predict the location of disease-associated SNPs. That disproportionately many pathogenic SNPs occur either at splice sites or at exon ends suggests that disruption of splicing is a key pathogenic process. Similarly, that genes with more exons (controlling for CDS length) are more likely to be disease-associated is consistent with a role for splicing (although this evidence is far from definitive owing to expression level covariance). What is striking, however, is that for our four other predictors, all report that pathogenic SNPs are most common where ESEs are least common: pathogenic SNPs are more common in 3′ gene domains and near short introns, they are associated with splice sites whose phase and nucleotide content are associated with low ESE density.

We might then predict an across-exon correlation between ESE density and density of pathogenic SNPs. To test this prediction we employ a set of ESE motifs with a low false positive rate but a high false negative rate (INT3 set) (Cáceres and Hurst 2013). For disease-associated genes, we calculate ESE density at exonic ends and correlate it with Dpi of these exons (see Materials and Method). A weak but significant negative correlation is found using Spearman's correlation ($\rho = -0.016$, $P = 0.009$) and Goodman–Kruskall gamma test ($\gamma = -0.023$, $P = 0.004$) (supplementary table S13, Supplementary Material online), the latter being less sensitive to multiple tied entries. This is consistent with the notion that pathogenic SNPs tend to be at ends of exons with low ESE density.

Discussion

One rationale for our findings is that many non-nonsense pathogenic SNPs disrupt splicing and that splice disruption is 1) more likely for mutations at exon ends and 2) more likely when a given exon has relatively few ESEs to provide robustness to the loss of any one. Thus, exon ends are hot spots of disease-associated mutations as these are where splice disrupting mutations occur. Likewise, genes with many potential alternative splice sites (those with many exons controlling for CDS length) are those more prone to be disease-associated. Within genes different parts are more ESE reinforced than others and an exon with a high density of ESEs might, for example, have multiple alternative ESE motifs to attract any given SR protein, should any given ESE mutagenically "fail." Conversely, some exons, notably those with low ESE density, are more prone to fail to splice correctly and these exons appear to be the hotspots for pathogenic SNPs. This we call the fragile exon model. Such a notion has precedent. For example, it has been observed that exons lacking a downstream intronic poly-G run, a known intronic splice enhancer, are more sensitive to 5′-splice site mutations (Lu et al. 2011). More generally, it will be informative to ask whether intronic SNPs that disrupt splicing (Kawase et al. 2007; Perriaud et al. 2014) are more likely to occur in proximity to low ESE density fragile exons.

The fragile exon model, however, need not be the only framework within which to interpret our data. As we are here analyzing pathogenic SNPs we must be alive to the notion that the population of SNPs is not random. It is possible, for example, that exons with more ESEs are as likely to be disrupted by SNPs (i.e., as fragile) but when such disruption occurs the effects are so dramatic (e.g., early embryonic lethality) that the SNP is never called "pathogenic." Indeed, were there exons whose correct splicing is so crucial, we might expect them to be supported by a very high density of ESEs to enable correct splicing in the absence of mutations. To differentiate between such models, one would need experimental evidence looking at the rate of missplicing in ESE-rich and ESE-poor exons within the same gene. If the rate is the same in the two then the second model might be more

parsimonious. If the rate of missplicing is higher in the low ESE class then the "fragile" exon model is more parsimonious.

We predict that 20–45% of SNPs are pathogenic owing to their effects on splicing. Interestingly, we see a similar trend for nonsense SNPs also to be enriched at exon ends (fig. 1*c*). This suggests that either nonsense mutations may exert their effects by mechanisms beyond introduction of a premature stop codon or that an aspect of our methodology to derive the above estimates may be incorrect. We assume that the excess of pathogenic SNPs at exon ends can be explained by the greater sensitivity of such domains to splice-altering mutations, as previously demonstrated (Woolfe et al. 2010). The extent of the excess then depends on what one considers the background rate. While our assumption appears well defended, we can nonetheless ask whether there might be other causes.

The process of nonsense-mediated decay (NMD) might also affect the rate of nonsense mutations that cause disease as a function of proximity to exon–exon junctions. This is because there is a gap of about 55 bp downstream of a premature stop codon that will not trigger NMD if there is an intron is located within this gap (Nagy and Maquat 1998). However, the end of exon enrichment for pathogenic SNPs is seen for non-nonsense SNPs as well as for nonsense ones, so rendering this an ungeneral explanation. A further possibility is that there might be a higher de novo mutation rate at exon ends for reasons unknown (association with nucleosomes might be conjectured [Kogan and Trifonov 2005; Chen et al. 2012]). This, however, appears unparsimonious as SNP rates at 4-fold degenerate sites not belonging to putative ESEs are if anything slightly lower at exon ends than cores (Cáceres and Hurst 2013). The fact that the rate appears very slightly lower might reflect a lower mutation rate (contra to what is required to explain the excess of end of exon pathogenic SNPs) or possibly reflects the imprecise definition of what constitutes an ESE. Where some true ESE hexamers left in the "non-ESE" class then they should be subject to the same purifying selection as witnessed for the true ESE hexamers. Given this uncertainty, direct parent-offspring sequencing to infer mutation rates as a function of intragene position would be a worthwhile enterprise.

Materials and Methods

Derivation of Exon and Intron Sequences from 30 Species

From "Table Browser" of UCSC (Karolchik et al. 2004) (http://genome.ucsc.edu/cgi-bin/hgTables, last accessed January 23, 2014) and FTP site of NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes, last accessed January 23, 2014), we obtained all available genes from 30 species (*Anolis carolinensis, Anopheles gambiae, Arabidopsis thaliana, Brachypodium distachyon, Caenorhabditis elegans, Callithrix jacchus, Cryptococcus neoformans, Dictyostelium discoideum, Drosophila melanogaster, Danio rerio, Ectocarpus siliculosus, Gallus gallus, Gorilla gorilla, Homo sapiens, Ictidomys tridecemlineatus, Meleagris gallopavo, Macaca mulatta, Mus musculus, Oryzias latipes, Oryza sativa, Pongo abelii, Plasmodium falciparum, Paramecium tetraurelia, Pan troglodytes, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Strongylocentrotus purpuratus, Sus scrofa, Takifugu rubripes, and Xenopus tropicalis*). Sequences without normal start codon (ATG) and stop codons (TAA, TAG, and TGA), the genes with internal stop codons, ambiguous nucleotides ("N"), and those without introns were all removed from the data set.

Collection of Information about Pathogenic SNPs

We downloaded disease-related information of sequence variation from Clinvar database (http://www.ncbi.nlm.nih.gov/clinvar/, last accessed November 24, 2014) (Landrum et al. 2014). In total, 9,818 pathogenic missense and silent mutations SNPs (which together we consider as non nonsense SNPs) were selected for further investigation (supplementary table S1, Supplementary Material online). From UCSC (http://genome.ucsc.edu/cgi-bin/hgTables, last accessed November 24, 2014), 1,747 gene sequences that contain pathogenic SNPs were derived (supplementary table S2, Supplementary Material online).

Comparison of Quantity of Exons between Disease Genes and Nondisease Genes

We compare the number of exons in disease-associated genes (1,747 genes) and nondisease genes (38,474 genes). To control for the effect of CDS length, we analyzed the residuals from loess regression of number of exons predicted by CDS size (supplementary table S2, Supplementary Material online). By Mann–Whitney *U* test analysis on the residuals of these two kinds of genes, we can assay whether number of exons in disease-associated genes is significant different from that of nondisease genes when CDS is controlled (supplementary table S2, Supplementary Material online).

Investigation of the Distribution of Pathogenic SNPs

By using our custom programs, we observed the distribution of pathogenic SNPs in exons or genes. The relevant information about these exons were also obtained, such as phase of exon ends, splice sites, distance from SNPs to the nearest splice sites, and flanking intron size (the flanking intron here is the nearest one to the SNP, if SNP is in the middle position, the flanking intron refers to the longer one; supplementary table S1, Supplementary Material online).

To test if pathogenic SNPs tend to be at end of exons, we observed their distribution in internal exons. Three domains ("$\leq$3 bp," "3–69 bp," and "$>$ 69 bp") were considered for enrichment of pathogenic SNPs. To establish the significance of the distributions, we summed up, by each SNP bearing exon, the length of each domain and divided them by total length of all the pathogenic exons separately thereby deriving the expected distribution (fig. 1; supplementary table S3, Supplementary Material online). We performed a similar enrichment analysis in order to exclude the effect coming from 5′ to 3′ distribution bias. We investigated the distribution of pathogenic missense and silent SNPs in 5′- and 3′-exon ends separately. A chi-squared test was used to examine the significance of all these results with fo/fe being employed to

detect whether the domain is preferred (> 1) or disliked (<1) (supplementary table S3, Supplementary Material online).

Estimation of the Proportion of Splice-Affecting Pathogenic SNPs at Exon Ends by the Background Removal Method

Assuming that mutations in exon cores never disrupt splicing, we defined the central 100 bp sequences (total length is "L") of internal exons longer than 300 bp as core regions and calculated the background nonsplicing related pathogenic SNP (total number is "*N*") rate by proportion (*N*/*L*) of pathogenic SNPs in this core region.

First, we consider 69 bp as exon end cutoff and calculate the total length of end regions in all internal exons (*M*). Then the predicted number of nonsplicing related SNPs is $P = M \times N/L$. After obtaining the number of all pathogenic SNPs at exon ends (Q), we can estimate the proportion of pathogenic SNPs that affect splicing at exon ends by $Ps = (Q - P)/(N + Q)$ (supplementary table S4, Supplementary Material online). If we consider that splice disrupting mutation in core occurs at an approximately 25% rate (Woolfe et al. 2010), the background nonsplicing related pathogenic SNP rate is $0.75 \times N/L$ (supplementary table S4, Supplementary Material online).

To examine whether the definition of exon end length affects robustness of this background removal method, we changed exon end cutoff from 50 to 100 bp by 5 bp every step and observed the change of proportion of splice-affecting pathogenic SNPs at exon ends (supplementary table S4, Supplementary Material online).

Biased Location of Pathogenic SNPs in 5′- and 3′-Halves of Genes

We observed the distribution of pathogenic SNPs in the absolute 5′-first half and the 3′-half (from ATG to stop) of each gene (supplementary table S1, Supplementary Material online). To consider the influence from the biased distribution of CDS sequence between first half and terminal half of genes, we performed a randomization. First, we randomly selected pseudo pathogenic SNPs in each disease-causing gene according to the real mutating nucleotides. Second, by calculating the ratio (expected values) of the number of pseudo SNPs at 3′-half to that at the 5′-half, we compared this expected ratio with that observed (ratio = 1.89). After repeating this comparison 100 times, we examine significance of this randomization process by $P = (n + 1)/(m + 1)$, where *n* is the number of expected values calculated after randomly selecting pseudo SNPs and meanwhile greater than (or equal to) the observed value and *m* is the number of randomization cycles 100 (supplementary table S5, Supplementary Material online). To consider the impact of mutations in the 3′-half of the CDS, we did analysis of the distribution of disease-causing SNPs in CDS sequence, testing for significance after Bonferroni correction (supplementary table S1, Supplementary Material online).

The Relationship between the Distribution of Pathogenic SNPs and Flanking Intron Size

An index of pathogenic SNP density (Dpi = number of pathogenic SNPs/exon length) was introduced to measure how easily a specific exon causes disease by pathogenic SNPs (supplementary table S6, Supplementary Material online). We performed Spearman's correlation analysis between Dpi and the log of flanking intron size only for disease SNP bearing internal exons. Here, flanking intron means the nearest intron. If the SNP happened to be in the middle then flanking intron is the longer one of the two equidistant introns.

Furthermore, to control for exon size and gene expression covariance we set up a simulation by randomly selecting pseudo SNPs according to the mutating nucleotide in internal exons (not including first and last exons) with the same gene. Then we determined the flanking intron size of the selected pseudo SNPs. By comparing the pseudo flanking intron size with the real one, we can determine how many randomly selected flanking introns are larger or smaller than the flanking introns of real pathogenic SNPs (result of each randomization simulation gives significance by Binomial Test). Repeating this trial 100 times allows estimation of *P* value from $p = (n + 1)/(m + 1)$, where *n* is the number of trials in which count of larger pseudo flanking intron size is greater than count of smaller ones and *m* is 100 (supplementary table S7, Supplementary Material online).

Additionally, because one exon is flanked by two introns on both sides, here, to consider Dpi value without any effect from flanking intron selection, we calculated Dpi values for the first half and the terminal half separately (5′-Dpi = number of pathogenic SNPs in 5′-half of exon/half length of exon, 3′-Dpi = number of pathogenic SNPs in 3′-half of exon/half length of exon). Genes ($N = 238$) that have at least five different internal exons containing pathogenic SNPs were selected to perform Spearman's correlation analysis between Dpi and flanking intron size. Based on these Dpi and flanking intron size of the 238 genes, we calculated ρ values from correlation analysis for each gene. Moreover, by Binomial distribution test, we can test if negative correlations tend to more common than expected by chance (supplementary table S8, Supplementary Material online).

Correlation between *cis*-Motif Usage with the Proportion of Phase Zero Splice Sites

For human genes, we calculated the number of splice sites in different phases (0, 1, 2) and the sum of numbers of all splice sites as a function of phase. A chi-squared test was used to examine if there is relation between coding phases and splice site usage (supplementary table S9, Supplementary Material online).

For comparative analysis we investigated all splice sites of 30 species, and classified them as "phase zero splice sites" (i.e.,

splice site in exons end with coding phase = 0) and "phase nonzero splice sites" (coding phase = 1 or phase = 2) (supplementary table S10, Supplementary Material online). Given that the composition of experimentally defined ESEs predicts the codons preferred near exon ends (Parmley and Hurst 2007; Cáceres and Hurst 2013), we presume that the frequency of distorted codon or amino acid usage in vicinity of exon junctions is a fair measure of *cis*-splice motif usage (Wu and Hurst 2015). To accord with an earlier analysis (Warnecke et al. 2008), the trend in usage of each codon and amino acid was investigated as a function of the distance from the exon–intron boundary up to a distance of 34 codons. The 5′- and 3′-ends were analyzed separately with the codon in direct proximity to the boundary being eliminated and the first and last exons being excluded. For each codon and amino acid under consideration, we determined, after Bonferonni correction, ρ and P value by two-tailed Spearman's correlation of proportional usage as a function of distance from the boundary. For each species we then calculated the proportion of codons or amino acids showing significant skew both at 5′- and 3′-ends across all exons and consider this the metric of *cis*-motif usage for that species. We calculated this metric by sampling 5,000 exons, so that there is no effect of the number of exons in different species, and a sample size uncorrected method, as we did in previous paper (Wu and Hurst 2015).

Based on the data set of 30 species genes, we also calculated the intronic parameter X (mean CDS length/gene length), which is an aggregate measure of intron size and density, for each species (supplementary table S10, Supplementary Material online). To allow for phylogenic nonindependence between data points, the program "Continuous" of BayesTraits (Pagel 1999) was used to study correlations between *cis*-motif usage and proportion of phase zero splice sites (supplementary table S10, Supplementary Material online) by a Markov chain Monte Carlo method. The phylogenetic tree with branch lengths is as previously employed (Wu and Hurst 2015). We abstracted the last harmonic mean from the result file, and took it as an estimation of marginal likelihood, to calculate the "Log BF" value and further test whether there is evidence for the correlation after phylogenetic correction (table 1).

To confirm the relationship between *cis*-motif usage and the proportion of phase zero splice site, we also correlated the proportion of phase zero splice site with three parameters: *X* (mean CDS length/gene length), *N* (introns per kb exon) and *M* (mean intron size) across species (table 2; supplementary table S10, Supplementary Material online).

Preference of Pathogenic SNPs for Splice Sites and Exon End Coding Phase

We explored the phase of exon ends and the tetranucleotide splice sites of the internal exons with pathogenic SNPs. Through randomly selecting a pseudo SNP (identical nucleotide with the mutating nucleotide) in each gene and scanning the exonic end phase and tetranucleotide splice sites of this pseudo SNP located exon, we test the significance of the real figures (observed value).

For all collected pseudo pathogenic SNPs, the numbers (Expected value) of exons with agGT, AGgt, and symmetric (0,0) exons were calculated separately. We repeated the randomization 100 times, and *P* value was obtained by the formula $p = (n + 1)/(m + 1)$, where n is the number of expected values calculated after randomly selecting pseudo SNPs and that are less than (or equal to) the observed value and m is 100 (the number of times of repeatedly performed) (supplementary table S11, Supplementary Material online).

We also do an assessment of the relative enrichment of pathogenic SNPs in (0,0) internal exons and 3′-AGgt splice site internal exons. The expected ratio is the length of all (0,0) (or 3′-AGgt splice site) internal exons divided by length of total internal exons. Similarly, the observed ratio is the number of pathogenic SNPs in (0,0) (or 3′-AGgt splice site) exons divided by total number of pathogenic SNPs. The metric of enrichment is "Observed ratio/Expected ratio" with significance being tested by a chi-squared Test (supplementary table S12, Supplementary Material online).

Correlation between Dpi and Exon End ESE Density in Disease-Associated Genes

To calculate ESE density, we employ an ESE candidate data set (INT3) composed of 84 *6-mer* (hexamer) motifs. This data set is likely to have a low false positive rate but a high false negative rate because it is an intersect of at least three well-identified ESE data sets (Cáceres and Hurst 2013). The list of INT3 hexamers may be obtained from supplementary table 1, Supplementary Material online, of the above article at http://www.genomebiology.com/content/supplementary/gb-2013-14-12-r143-s1.xlsx. Based on this INT3 data set, we calculate ESE densities of each internal exon end and, for exons longer than 138 (69×2) bp, take mean ESE density value of two 69 bp ends at both sides of exon (if exon is shorter than 138 bp, then the whole exon are regarded as end region) for correlation analysis with Dpi. Significance of the correlation is examined by Spearman's correlation and Goodman–Kruskall gamma test (supplementary table S13, Supplementary Material online, if there is no pathogenic SNP in a exon, Dpi of this exon = 0).

Supplementary Material

Supplementary tables S1–S13 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a University Research Studentship from the University of Bath to X.W., and Medical Research Grant MR/L007215/1 and European Research Council (ERC) grant ERC-2014-ADG 669207 to L.D.H.

References

Amendt BA, Si ZH, Stoltzfus CM. 1995. Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. *Mol Cell Biol.* 15:6480.

Bali V, Bebok Z. 2015. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int J Biochem Cell Biol.* 64:58–74.
Berget SM. 1995. Exon recognition in vertebrate splicing. *J Biol Chem.* 270:2411–2414.
Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* 25:106–110.
Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* 14:R143.
Caputi M, Kendzior RJ Jr, Beemon KL. 2002. A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer. *Genes Dev.* 16:1754–1759.
Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol.* 62:89–98.
Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 3:285–298.
Cartegni L, Krainer AR. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet.* 30:377–384.
Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7:98–108.
Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. 2012. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* 335:1235–1238.
Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.
Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452:423–428.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007–1013.
Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev.* 17:419–437.
Fedorov A, Suboch G, Bujakov M, Fedorova L. 1992. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res.* 20:2553–2557.
Gavrilov DK, Shi X, Das K, Gilliam TC, Wang CH. 1998. Differential SMN2 expression associated with SMA severity. *Nat. Genet.* 20:230–231.
Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* 6:1197–1211.
Graveley BR, Hertel KJ, Maniatis T. 1998. A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J.* 17:6747–6756.
Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. *Trends Genet.* 30:308–321.
Kan JL, Green MR. 1999. Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. *Genes Dev.* 13:462–471.
Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32:D493–D496.
Kawase T, Akatsuka Y, Torikai H, Morishima S, Oka A, Tsujimura A, Miyazaki M, Tsujimura K, Miyamura K, Ogawa S, et al. 2007. Alternative splicing due to an intronic SNP in HMSD generates a novel minor histocompatibility antigen. *Blood* 110:1055–1063.
Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360–1374.
Kogan S, Trifonov EN. 2005. Gene splice sites correlate with nucleosome positions. *Gene* 352:57–62.
Krawczak M, Reiss J, Cooper DN. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet.* 90:41–54.
Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42:D980–D985.
Lavigueur A, La Branche H, Kornblihtt AR, Chabot B. 1993. A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. *Genes Dev.* 7:2405–2417.
Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A.* 98:11193–11198.
Long M, Rosenberg C, Gilbert W. 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc Natl Acad Sci U S A.* 92:12495–12499.
Lu ZX, Jiang P, Cai JJ, Xing Y. 2011. Context-dependent robustness to 5′ splice site polymorphisms in human populations. *Hum Mol Genet.* 20:1084–1096.
Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827–1836.
Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Mooney SD. 2014. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* 15:R19.
Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci.* 23:198–199.
Nelson KK, Green MR. 1988. Splice site selection and ribonucleoprotein complex assembly during in vitro pre-mRNA splicing. *Genes Dev.* 2:319–329.
Nissim-Rafinia M, Kerem B. 2002. Splicing regulation as a potential genetic modifier. *Trends Genet.* 18:123–127.
Pagani F, Raponi M, Baralle FE. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A.* 102:6368–6372.
Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol.* 24:1600–1603.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14.
Perriaud L, Marcel V, Sagne C, Favaudon V, Guedin A, De Rache A, Guetta C, Hamon F, Teulade-Fichou MP, Hainaut P, et al. 2014. Impact of G-quadruplex structures and intronic polymorphisms rs17878362 and rs1642785 on basal and ionizing radiation-induced expression of alternative p53 transcripts. *Carcinogenesis* 35:2706–2715.
Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet.* 24:590–594.
Rowen L, Young J, Birditt B, Kaur A, Madan A, Philipps DL, Qin S, Minx P, Wilson RK, Hood L, et al. 2002. Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* 79:587–597.
Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD. 2005. Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol* 60:99–104.
Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet.* 12:683–691.
Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14:802–813.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9:R29.

Wirth B, Brichta L, Hahnen E. 2006. Spinal muscular atrophy: from gene to therapy. *Semin Pediatr Neurol.* 13:121–131.

Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. *Genome Biol.* 11:R20.

Wu X, Hurst LD. 2015. Why selection might be stronger when populations are small: intron size and density predict within and between-species usage of exonic splice associated cis-motifs. *Mol Biol Evol.* 32:1847–1861.

Zheng ZM, Quintero J, Reid ES, Gocke C, Baker CC. 2000. Optimization of a weak 3′ splice site counteracts the function of a bovine papillomavirus type 1 exonic splicing suppressor in vitro and in vivo. *J Virol.* 74:5902–5910..

Chapter V.

Are there tissue-specific ESEs?

Contributions

All analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst.

Are there tissue-specific ESEs?

Abstract

In this chapter, some preliminary analysis as regards whether there are tissue-specific ESEs show encouraging results for transgenesis and gene therapy. Based on comparison of ESE usage patterns, codon usage bias and expression levels between tissue-specific genes, I found ESE usage is not a function of expression level. Furthermore, for these tissue-specific genes, no different ESE usage pattern has been detected between disease associated genes and non-disease ones. Interestingly, genes with higher ESE density tend to express in more tissues, which indicates tissue-specific genes should have few ESEs, meaning it might be possible to delete some ESEs with little effect. While this potential application value in practice still needs further scrutiny or even requires experimental confirmation, ESEs, SR proteins and special exons are all worth of investigating in tissue-specific context.

Introduction

Previous studies suggest that codons in exon end regions often have at least two functions: not simply encoding for proteins in translation but also controlling splicing post-transcriptionally. For coding, 61 codons are corresponding to amino acids. Of the 20 amino acids, 18 have more than one codon specifying them. Commonly these synonymous codons are used unequally, this being codon usage bias (Ikemura 1981). Different species show different trends in codon usage bias. The same is also true when comparing genes expressed uniquely in one tissue, that is we see between-tissue within-species differences in codon usage, although the effect is weak (Plotkin, Robins, and Levine 2004; Sémon, Lobry, and Duret 2006). This between-tissue difference in codon usage prompted us to ask whether there might also be a difference in *cis*-motifs usage between tissue-specific genes.

In this final chapter I perform some preliminary and exploratory analysis of ESE usage in a tissue-specific context. In particular I ask whether, looking at genes expressed in only one tissue, I observe different ESEs usage patterns in different tissues. This means both whether certain ESEs are used more in certain tissues and whether ESE density varies across genes expressed in only one tissue as a function of the tissue. The main concerns here are biotechnological applications rather than evolutionary insights. That is to say, in

mouse transgenesis and in human gene therapy, the aim is commonly to replace a non-functioning gene with a functioning one. However, as the wild-type gene is often too large to be successfully integrated, it is commonplace to remove all but the first intron (for reasons unknown the first intron's presence promotes expression) (Klamut et al. 1996; Shen et al. 1997; Clancy and Hannah 2002). A challenge is to make sure the transgene is expressed in the correct tissue and only in the correct tissue. Use of tissue-specific promoters is here helpful, but might use of tissue specific ESEs also aid the correct splicing of the gene only in the tissue concerned? If we could identify tissue specific ESEs we could in principle reinforce the first intron with tissue specific ESEs by modification of silent sites. The net aim would be to promote correct splicing in the focal tissue and simultaneously to encourage incorrect splicing (including non-splicing) in undesirable tissues. Incorrect splicing in a different tissue would likely lead to initiation of the nonsense mediated decay pathway, so minimizing the deleterious effect of spurious expression.

In principle then we could engineer in tissue specific ESEs and engineer out ESEs that might promote splicing in the undesired tissues. Both can be done via intelligent modification of synonymous sites. However, we need also to check for covariates. Is it for example the case that ESE usage varies with expression level – might a highly expressed gene, for example, need more ESEs, in which case deletion of ESEs by changing synonymous sites might be foolhardy. Likewise, does ESE density vary by tissue? If it does this might imply again caution against deletion of ESEs in some genes expressed in certain tissues. We might a priori expect there to be such differences as a) SR proteins show tissue-specific patterns of expression (Jumaa, Guenet, and Nielsen 1997; Liu, Zhang, and Krainer 1998), and b) different SR proteins work through different sequences (Liu, Zhang, and Krainer 1998). Thus, one of my aims is to look for ESEs with exceptional or unusual usage patterns across tissue-specific genes, as a function of the tissue.

Results

1. Highly expressed tissue specific genes have the same ESE density as others

As noted above it would be instructive to know whether gene expression level predicts ESE usage as expression level is a key issue for transgene engineering (Emilsson et al. 2008). Here, we define genes are "tissue-specific" due to their dramatically high expression in only one tissue (at least five times higher than the expression in all other tissues) (Uhlen et al. 2015). Using this definition we define a relatively small set of genes

(indeed much smaller than a much looser definition employed by others (Sémon, Lobry, and Duret 2006)). Is there then a correlation between ESE density and expression levels in these tissue specific genes? We employ a strong consensus ESE set (INT3) and a larger sample embracing both this consensus set and the RESCUE motifs, these being shown to be prototypical ESEs (Cáceres and Hurst 2013). We find that for neither INT3 nor INT3+RESCUE datasets is there a correlation between density and expression level (Supplementary table S2). This suggests that modification of ESE density is unlikely to modify or be modified by expression level. This is helpful when thinking about deleting ESEs via synonymous site modification.

By contrast, based on codon usage preference of ribosomal protein encoding genes, we do find there are significant positive correlation between whole CDS codon usage bias (CAI) (Sharp and Li 1987) and expression (rho = 0.19, $P = 2.06 \times 10^{-14}$) of tissue-specific genes. This suggests that highly expressed genes might be under selection to use optimal codons, although this finding is contentious in the human context (Francino and Ochman 1999; Urrutia and Hurst 2001; Galtier 2003) and control for isochore effects (by far the predictor of codon usage profiles in human genes) would be needed to make any strong inference. In addition we observe a negative correlation between ESE density and CAI (INT3: rho = -0.13, $P = 1.74 \times 10^{-7}$, INT3+RESCUE: rho = -0.38, $P = 2.94 \times 10^{-57}$) (Supplementary table S2). This suggests that, for tissue-specific genes, codon usage bias of exon core region might predict expression level; and ESE candidates, *6-mer* nucleotides sequence which are enriched at exon ends, are not used in same manner. Analysis of this issue, including isochore control, is left to future study. Here the main conclusion I wish to draw is that ESE usage at exon ends is unaffected by expression level, even if alternative metrics of codon usage do so correlate. As a consequence deletion of ESEs in transgenes is unlikely to modify expression level.

2. There are differences ESE usage, codon usage and expression levels between tissues

Previous studies suggest different usage patterns of codon usage between tissues, although there is a debate about the scale and the analysis method (Plotkin, Robins, and Levine 2004; Sémon, Lobry, and Duret 2006). One possible reason for this could be relate to the usage of tissue specific ESEs which in turn distort codon usage patterns. Alternatively we may be witnessing effects mediated by some other mode of evolution of codon usage (possibly a function of expression level). Here then I ask whether there are differences in ESE usage, codon usage and expression levels between tissues. I find that ESE density,

CAI and expression level all show significant between-tissue heterogeneity (From ANOVA: Supplementary table S3). I infer that different genes expressed in certain tissues uses ESEs more or less than those employed in other tissues. As there is no correlation between expression level and ESE usage, the heterogeneity in expression level, whilst evident, is not an issue as regards engineering tissue specific genes. The CAI variation may in part be explained by expression level variation, but again this will need to have isochore effects controlled for.

3. Non-homogeneous usage of *6-mer* ESE candidates across different tissues

That ESE density varies between tissues is instructive as regards our aim to better design transgenes. But the key issue is whether certain *6-mer* ESE candidates are used more in certain tissues than others i.e. are some ESEs effectively tissue specific? Were this true then we could engineer these in or out.

In each tissue, I calculated the density of each *6-mer* ESE candidate at exon ends based on INT3 dataset. To avoid effects of ESEs overlapping, two types of density of each ESE were employed, "number of ESEs/length of exon ends" and "number of nucleotides in ESEs/length of exon ends". Some non-homogeneous usage is indeed detected across different tissues (Supplementary table S4). These are best visualized as heatmaps (Fig. 1). From this we infer that there are indeed some ESEs that seem to have striking patterns of relative tissue specificity.

To express this numerically I employ the entropy measure tau, used before to define tissue specificity in expression (see e.g. (Weber and Hurst 2011)). Tau is defined as:

$$\tau = \frac{\sum_{j=1}^{n}(1 - \frac{S_{(j)}}{S_{(max)}})}{n-1}$$

where $S(j)$ is the density of the ESE in genes expressed uniquely in tissue j, S_{max} is the maximum density for that ESE across all tissues and n is the number of tissues. Moreover, this metric affords us the possibility to check for significance of the tissue specificity. If we randomize the numbers across the matrix, we can ask for any given tau value, how often we would expect to see tau as big or bigger than that actually reported for any given ESE. These results and corresponding tau values are presented in Table 2. From significance test, "AAGAAT", "AACCAG", and "AACTGG" show notable tissue specificity. This comes with a severe caveat that none of the significance levels passes multitest correction.

4. Genes that express in more tissues have higher ESE density

A possible corollary of the above finding of relatively tissue specific ESEs, is that genes expressed in many tissues might need a richer palette of ESEs and hence a higher density of ESEs. Is this true? To test for such a correlation I established a dataset of human genes and determined the number of tissues in which they were expressed (out of 27 tissues) (Supplementary table S5). I find that ESEs densities of the genes (n=15326), which have at least two introns, correlate positively and significantly with the proportion of tissues in which this gene expressed (rho = 0.06, $P = 4.40 \times 10^{-14}$). This indicates genes with higher ESE density tend to express in more tissues, therefore tissue-specific genes should have few ESEs. Tissue specific genes thus have a lower density of ESEs, which also suggests one might be able to ablate ESEs of tissue specific genes with little harm. We note however the effect is weak, and the interpretation far from clear, as there may be covariance with intron size and density.

5. There is relationship of tissue-specific genes with pathogenic SNPs

Gene therapy is typically only needed to replace disease-causing genes with the wild-type version. Above I have presumed that disease-associated and unassociated tissues specific genes are identical as regards their ESE behavior. To address this issue I cross-reference the 1747 genes with pathogenic SNPs identified in the previous chapter (Chapter IV) with tissue specific genes. By matching the IDs with the tissue-specific genes, we can calculate the proportion of tissue-specific genes with pathogenic SNPs in different tissues (Supplementary table S6). I then ask whether disease and non-disease tissue specific genes differ in any important parameters.

I find that ESE density (using INT3) is not different (Mann Whitney U test: $P = 0.25$), while CAI (Mann Whitney U test: $P = 1.78 \times 10^{-6}$) and expression levels (Mann Whitney U test: $P = 1.99 \times 10^{-8}$) are different between disease and non-disease tissue-specific genes (Supplementary table S6) (Fig. 2). I conclude that ESE density between disease and non-disease tissue-specific genes is no different but codon usage bias and expression levels are all likely to be higher in disease-associated tissue-specific genes. I conclude that conclusions as regards ESEs in disease associated tissue specific genes can probably be fairly extrapolated from results for tissue specific genes more generally.

Discussion

As regards the prospects of altering ESE content to manipulate transgenes the results of this brief analysis are largely encouraging. There are such things are relatively tissue specific ESEs (prior to multi-test control) and ESE usage is not a function of expression level. Disease associated genes appear to behave the same as regards ESEs as non-disease genes. Moreover, I found genes with higher ESE density tend to express in more tissues. This indicates tissue-specific genes should have few ESEs, meaning it might be possible to delete some ESEs with little effect. Whether in practice these insights prove helpful will require experimental confirmation. I note that not all tissues have an ESE particular to that tissue.

Materials and Methods

ESEs usage of tissue-specific genes

We download tissue-specific gene information (tissue type, gene ID and expression level) from "The human protein atlas" (http://www.proteinatlas.org/). These genes are defined as "tissue-specific" (or "tissue enriched") due to their mRNA levels in a particular tissue being at least five times those in all other tissues (Uhlen et al. 2015). By gene IDs list and our perl program, we obtain reference sequences (hg38) of tissue enriched genes in which there are at least 2 introns (we do this for analysis on ESE density of internal exons). Only those (n=1595) expressed in the tissues (n=19) that have more than 5 tissue-specific genes are collected for further analysis (Supplementary table S1) (Table 1).

We obtain exon ends (if an exon is longer than 138 bp, both sides of 69 bp exon ends are used; otherwise the whole exon are regarded as "end region") ESE density by calculating the proportion of nucleotides involved in ESEs candidate (INT3) in length of exon end region. Then ESEs density of a gene is the average of ESEs densities of its all internal exons.

Codon usage bias of tissue-specific genes

We use CAI (Codon Adaptation Index) to measure the codon usage patterns of tissue-specific genes. First we set up a reference dataset of 155 human ribosomal protein encoding genes (From: RPG- the Ribosomal Protein Gene database) (Nakao, Yoshihama, and Kenmochi 2004), and then by "Countcodon program" (http://www.kazusa.or.jp/codon/countcodon.html) establish a reference codon table. CAI values can be calculated by a program "CAIcal" (CAIcal_ECAI_v1.4.pl) (Supplementary table S1).

Correlation analysis on ESE usage, codon usage and expression levels

By Spearman's correlation analysis, we do pairwise correlation of ESEs usage (ESEs density based on "INT3" and "INT3+RESCUE" candidate datasets), codon usage (CAI) and expression levels (FPKM) (Supplementary table S2).

Tissue specificity analysis on ESE usage, codon usage and expression levels

We do a test of "homogeneity of variance" in R for ESE density, CAI and expression levels separately. All of them are suitable for ANOVA analysis to see whether there are any significant differences between independent groups (different tissues) (Supplementary table S3).

Usage of *6-mer* ESE candidates in different tissues

In each tissue, we calculate density of each *6-mer* ESE candidates based on INT3 dataset. Because we don't know if ESEs overlapping affect the result, two type of density are observed, one is "number of ESEs/length of exon ends"; the other one is "number of nucleotides in ESEs/ length of exon ends". Density of *6-mer* ESE candidates for a tissue comes from average of its tissue-specific genes ESE density, and genes ESE candidate density is from average of exons ESE candidate density. According to these data in different tissues (Supplementary table S4), we make heatmaps to show different usage patterns.

To compare usage pattern of *6-mer* ESE candidates across tissues in a numerically manner, I employ an index similar to the entropy measure tau, used before to define tissue specificity in expression (see e.g. (Weber and Hurst 2011)). Tau is here defined as:

$$\tau = \frac{\sum_{j=1}^{n}(1 - \frac{S_{(j)}}{S_{(max)}})}{n-1}$$

where $S(j)$ is the density of the ESE in genes expressed uniquely in tissue j, S_{max} is the maximum density for that ESE across all tissues and n is the number of tissues. Moreover, I check for significance of the tissue specificity by randomizing the numbers across the matrix. For any given tau value we repeat 1000 times (n) to set up a pseudo usage pattern across tissues, and then ask for, how often we would expect to see tau as big or bigger than that actually reported one (m). So, by p=(m+1)/(n+1), I can calculate P value for each ESE candidate (Table 2).

Relationship between ESEs usage and breadth of expression in tissues

From database "The human protein atlas" (http://www.proteinatlas.org/), we download dataset (mcp.M113.035600-2) of human genes expression level in 27 tissues (Supplementary table S5). Also, ESEs densities of the genes (n=15326), which have at least two introns, are calculated and correlated with the proportion of tissues in which this gene expressed (Spearman's correlation analysis).

Observation of the tissue-specific genes with pathogenic SNPs in different tissues

In previous study, we defined 1747 genes with pathogenic SNPs according to experimental data. Then matching the IDs with tissue-specific genes in this research, we can find the proportion of tissue-specific genes with pathogenic SNPs in different tissues (Supplementary table S6). We also do boxplots to show difference of ESE density, CAI and expression levels between disease and non-disease tissue-specific genes (Supplementary table S6).

Tables

Table 1. Tissues and their number of tissue-specific genes

No.	Tissue	Number of tissue-specific genes
1	adipose tissue	17
2	adrenal gland	30
3	bone marrow	60
4	cerebral cortex	249
5	esophagus	34
6	fallopian tube	46
7	heart muscle	27
8	kidney	51
9	liver	152
10	lung	17
11	pancreas	38
12	placenta	64
13	prostate	16
14	salivary gland	33
15	skeletal muscle	89
16	skin	60
17	stomach	20
18	testis	573
19	thyroid gland	19

Table 2. Non-homogeneous usage of *6-mer* ESE candidates across different tissues

Type 1 ESE density[a]			Type 2 ESE density[b]		
ESE_6_mer	Tau	P_value	ESE_6_mer	Tau	P_value
AAGAAT	0.8956	0.0100	AAGAAT	0.8956	0.0110
AACCAG	0.8864	0.0200	AACCAG	0.8864	0.0180
AACTGG	0.8615	0.0619	AACTGG	0.8615	0.0470
GGAAGA	0.8512	0.0739	GGAAGA	0.8512	0.0589
GAGGAA	0.8340	0.0929	GAGGAA	0.8340	0.0889
GAAGTA	0.8301	0.0949	GAAGTA	0.8301	0.0929
AGAAGT	0.8305	0.0949	AGAAGT	0.8305	0.0929
AACAAC	0.8179	0.1069	GTTGGA	0.8165	0.1149
GTTGGA	0.8164	0.1089	AACAAC	0.8168	0.1149
GAAGAC	0.8139	0.1159	GAAGAC	0.8139	0.1209
GACATC	0.8060	0.1299	GACATC	0.8060	0.1309
GGAGGA	0.8021	0.1339	CCTGGA	0.7965	0.1538
CCTGGA	0.7965	0.1419	GGAGGA	0.7959	0.1538
ACTGGA	0.7933	0.1518	ACTGGA	0.7933	0.1588
GACGAA	0.7867	0.1668	GACGAA	0.7867	0.1768
AGACGA	0.7848	0.1728	AGACGA	0.7848	0.1868
AAGAAC	0.7762	0.1898	AAGAAC	0.7762	0.2048
GAAGGA	0.7749	0.1918	GAAGGA	0.7746	0.2108
AGAAAC	0.7551	0.2458	AGAAAC	0.7551	0.2727
GACCTG	0.7472	0.2667	GACCTG	0.7472	0.2997
GAAGAA	0.7439	0.2757	GAAGAA	0.7370	0.3287
GAAGAG	0.7354	0.3077	GAAGAG	0.7352	0.3347
TGAAGA	0.7343	0.3117	TGAAGA	0.7343	0.3417
TCCTGG	0.7288	0.3247	TCCTGG	0.7288	0.3556
AACTAC	0.7124	0.3846	AACTAC	0.7124	0.4026
CGAGGA	0.7076	0.4076	CGAGGA	0.7077	0.4206
AGTGAC	0.7019	0.4386	AGTGAC	0.7019	0.4476
TCAAGA	0.7006	0.4416	TCAAGA	0.7006	0.4525
TACCTG	0.6947	0.4665	TACCTG	0.6947	0.4745
GCAGAA	0.6940	0.4705	GCAGAA	0.6940	0.4775
ATTGGA	0.6884	0.4935	ATTGGA	0.6884	0.4945
AACTTC	0.6858	0.5025	AACTTC	0.6858	0.4985
GAGGAT	0.6795	0.5285	GAGGAT	0.6795	0.5235
CGAAGA	0.6643	0.5804	CGAAGA	0.6643	0.5924
AGAGAA	0.6577	0.5984	AGAGAA	0.6577	0.6184
CTGAAG	0.6565	0.6024	CTGAAG	0.6565	0.6194
CAGAAG	0.6463	0.6444	CAGAAG	0.6463	0.6523
ATCTGC	0.6462	0.6464	ATCTGC	0.6462	0.6533
GACTTC	0.6454	0.6494	GACTTC	0.6454	0.6563
AACAGA	0.6370	0.6793	AACAGA	0.6370	0.6883
AAGAAA	0.6332	0.6883	AAGAAA	0.6327	0.7003
GAGGAG	0.6262	0.7203	AATGAC	0.6261	0.7193
AATGAC	0.6261	0.7213	AGGAAC	0.6205	0.7313
AGGAAC	0.6205	0.7353	AACCTG	0.6199	0.7333
AACCTG	0.6199	0.7373	AAGGAA	0.6175	0.7433
AAGGAA	0.6178	0.7493	TGAGAA	0.6150	0.7493
TGAGAA	0.6150	0.7582	GAGGAG	0.6130	0.7552
CAAAGA	0.6123	0.7692	CAAAGA	0.6123	0.7572
GAAAGA	0.6005	0.8072	TGAAGG	0.6005	0.7902
TGAAGG	0.6004	0.8072	GAAAGA	0.6000	0.7912
GAAACT	0.5982	0.8152	GAAACT	0.5982	0.7962
GAAGTT	0.5954	0.8262	GAAGTT	0.5954	0.8032
ACAGAA	0.5942	0.8302	ACAGAA	0.5942	0.8082
GAAGCA	0.5933	0.8342	GGAGAT	0.5927	0.8112
GGAGAT	0.5927	0.8352	GAAGCA	0.5933	0.8112
GAAGAT	0.5904	0.8402	GAAGAT	0.5904	0.8182
TGAAGC	0.5867	0.8511	TGAAGC	0.5867	0.8322
AGAAGC	0.5836	0.8541	AAGGAC	0.5839	0.8392
AAGGAC	0.5839	0.8541	AGAAGC	0.5836	0.8422
AAGAGA	0.5827	0.8551	AAGAGA	0.5823	0.8462
AGAGGA	0.5759	0.8651	AAGAAG	0.5765	0.8551
GCAAGA	0.5727	0.8761	AGAGGA	0.5759	0.8561
AGAAGA	0.5726	0.8761	GCAAGA	0.5727	0.8661

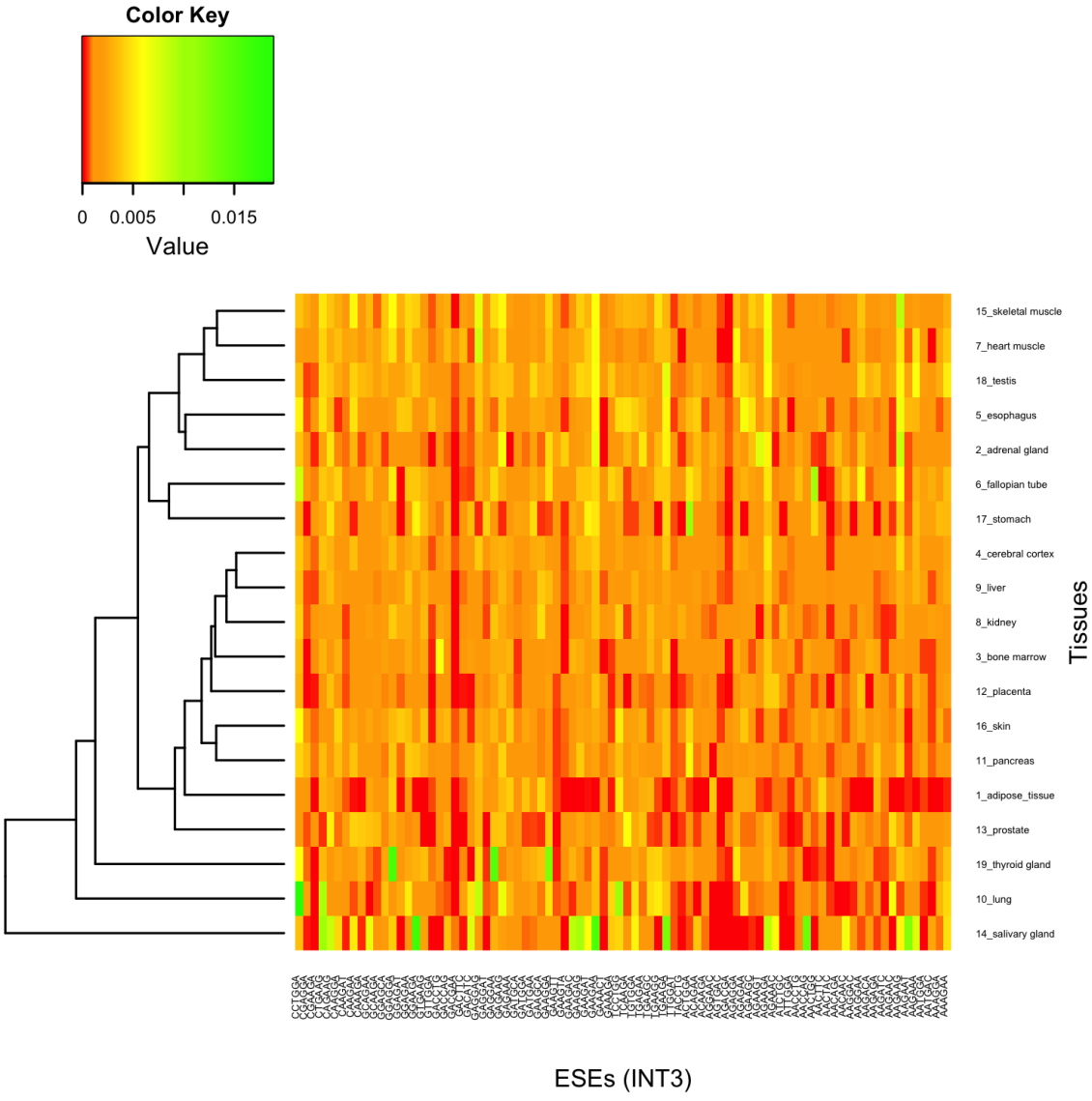
TGTGGA	0.5693	0.8831	TGTGGA	0.5693	0.8781
GATGGA	0.5675	0.8891	AGAAGA	0.5676	0.8841
CAAGAA	0.5586	0.9101	GATGGA	0.5659	0.8891
TTGGAT	0.5540	0.9191	CAAGAA	0.5586	0.9061
AAGAAG	0.5535	0.9211	TTGGAT	0.5541	0.9101
GATGAA	0.5404	0.9421	GATGAA	0.5404	0.9251
AAGACA	0.5382	0.9451	AAGACA	0.5382	0.9281
GAGAAG	0.5258	0.9570	GAGAAG	0.5257	0.9451
CAAGAT	0.5222	0.9640	CAAGAT	0.5222	0.9481
AAAGAA	0.5152	0.9690	AAAGAA	0.5172	0.9550
GAGAAA	0.4961	0.9790	GAGAAA	0.4961	0.9710
GATGCA	0.4905	0.9800	GATGCA	0.4905	0.9770
GACCAG	0.4787	0.9870	GACCAG	0.4787	0.9820
AAAGGA	0.4788	0.9870	AAAGGA	0.4788	0.9820
GGAGCA	0.4708	0.9910	GGAGCA	0.4708	0.9840
CAAGGA	0.4668	0.9920	AATGGA	0.4682	0.9840
GGAGAA	0.4644	0.9920	CAAGGA	0.4668	0.9850
AAGATC	0.4557	0.9920	GGAGAA	0.4644	0.9860
AATGGA	0.4682	0.9920	AAGATC	0.4557	0.9910
GTGAAG	0.3950	0.9980	ACAAGA	0.4176	0.9980
ACAAGA	0.4167	0.9980	GTGAAG	0.3950	1.0000

[a]: ESE density = number of nucleotides in ESEs/length of exon ends
[b]: ESE density = number of ESEs/length of exon ends

Figures

Fig. 1 heatmaps to show different usage patterns of *6-mer* ESE candidates

(a) heatmap showing different usage patterns of *6-mer* ESE candidates (number of nucleotides in ESEs/ length of exon ends)



(b) heatmap showing different usage patterns of *6-mer* ESE candidates (number of ESEs/length of exon ends)

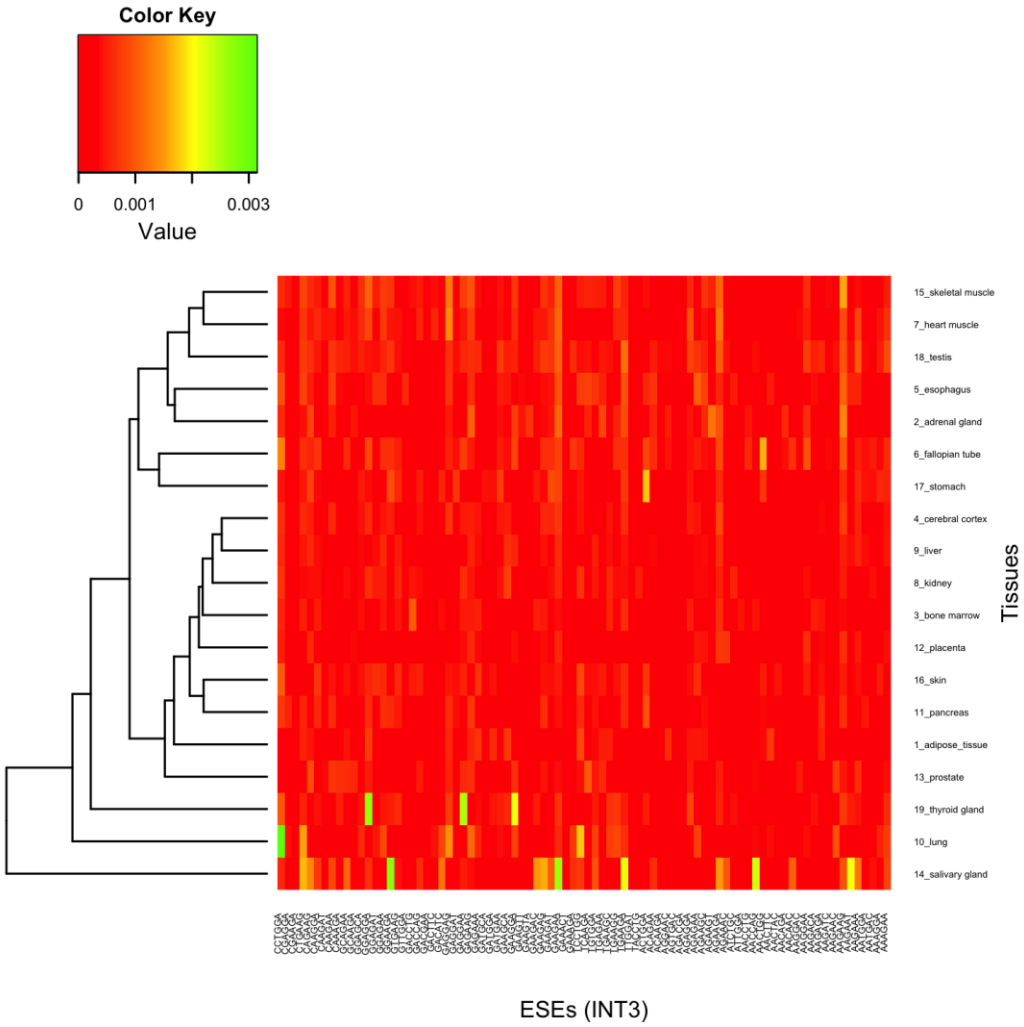
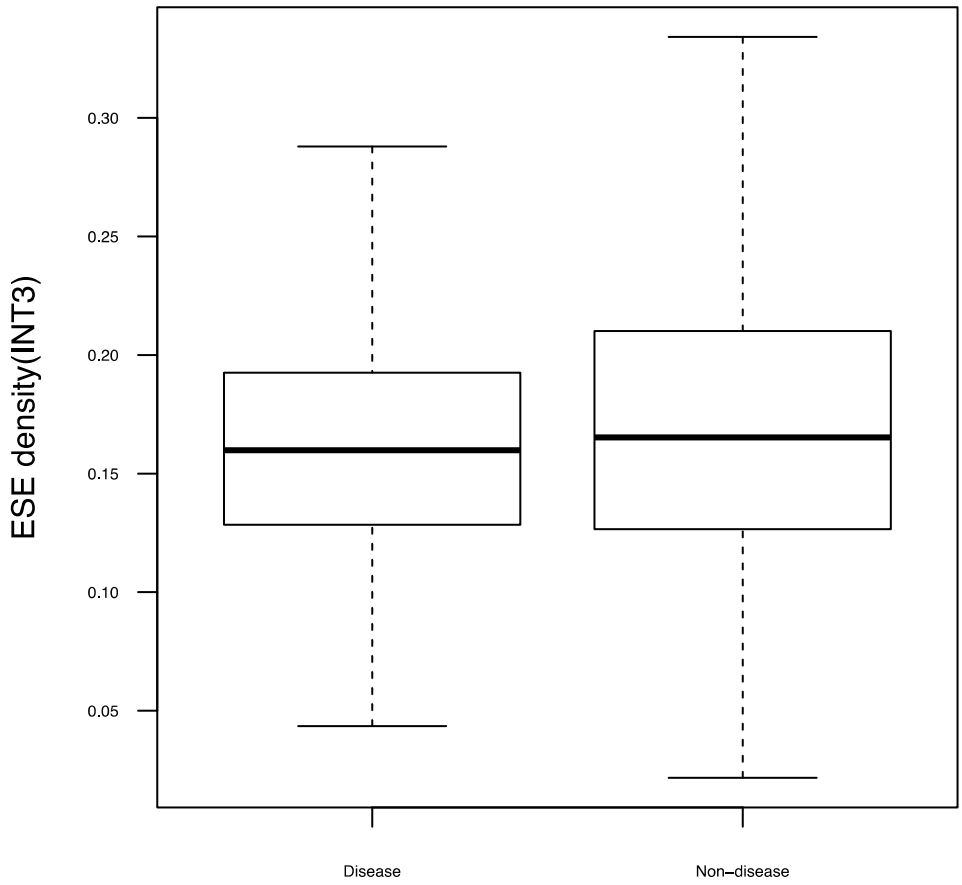
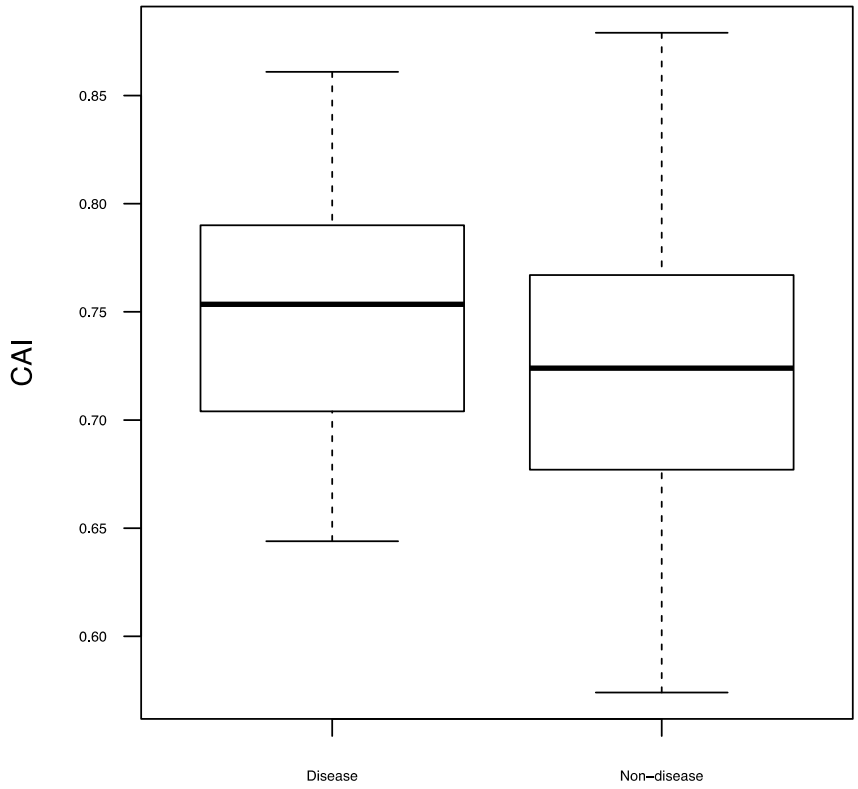


Fig. 2 difference of ESE density, CAI and expression levels between disease and non-disease tissue-specific genes

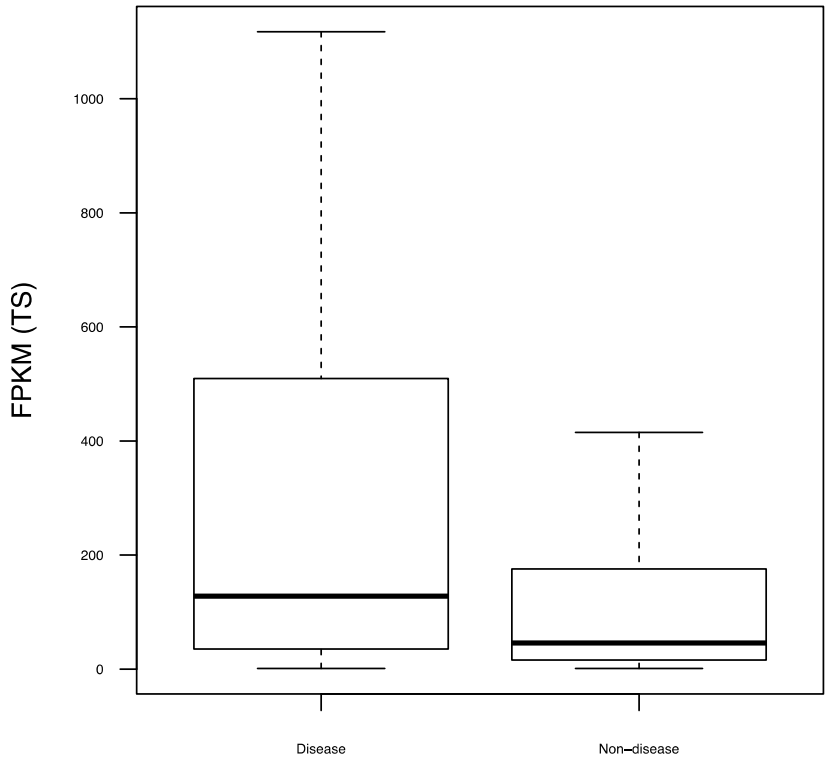
(a) Difference of ESE density between disease and non-disease tissue-specific genes



(b) Difference of CAI between disease and non-disease tissue-specific genes



(c) Difference of expression levels between disease and non-disease tissue-specific genes



Reference

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Clancy M, Hannah LC. 2002. Splicing of the maize Sh1 first intron is essential for enhancement of gene expression, and a T-rich motif increases expression without affecting splicing. *Plant Physiol* 130:918-929.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452:423-428.

Francino MP, Ochman H. 1999. Isochores result from mutation not selection. *Nature* 400:30-31.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet* 19:65-68.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 151:389-409.

Jumaa H, Guenet JL, Nielsen PJ. 1997. Regulated expression and RNA processing of transcripts from the Srp20 splicing factor gene during the cell cycle. *Mol Cell Biol* 17:3116-3124.

Klamut HJ, Bosnoyan-Collins LO, Worton RG, Ray PN, Davis HL. 1996. Identification of a transcriptional enhancer within muscle intron 1 of the human dystrophin gene. *Hum Mol Genet* 5:1599-1606.

Liu HX, Zhang M, Krainer AR. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12:1998-2012.

Nakao A, Yoshihama M, Kenmochi N. 2004. RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* 32:D168-170.

Plotkin JB, Robins H, Levine AJ. 2004. Tissue-specific codon usage and the expression of human genes. *Proc Natl Acad Sci U S A* 101:12588-12591.

Sémon M, Lobry JR, Duret L. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol Biol Evol* 23:523-529.

Sharp PM, Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281-1295.

Shen Y, Liu J, Wang X, Cheng X, Wang Y, Wu N. 1997. Essential role of the first intron in the transcription of hsp90beta gene. *FEBS Lett* 413:92-98.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419.

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159:1191-1199.

Weber CC, Hurst LD. 2011. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. *Genome Biol* 12:R23.

Chapter VI. Discussion

While it is tempting to suppose that the evolution of a gene is largely dictated by selection on the protein product, it is now clear that much selection is also involved in the manufacture of the product (Chamary and Hurst 2005; Carlini and Genut 2006; Hurst 2006; Nackley et al. 2006; Parmley, Chamary, and Hurst 2006; Parmley et al. 2007; Warnecke, Weber, and Hurst 2009; Bartoszewski et al. 2010; Brest et al. 2011; Cáceres and Hurst 2013; Gartner et al. 2013; Lawrie et al. 2013). At least in humans (or mammals more generally) splice-related constraints feature strongly in this manufacture side of the equation. But are all species with introns equally affected? Are all genes or exons equally affected? If we can predict which exons have more splice support can we also predict which exons, and where in exons, disease associated mutations occur? In my thesis, I address such issues using both intra and inter-specific analysis.

In my first analysis I follow up on the suggestion, from both experimental and comparative analysis, that large introns are harder to splice accurately. For instance, splice rates decrease if extra sequence is experimentally inserted into introns (Klinz and Gallwitz 1985; Luehrsen and Walbot 1992; Fox-Walsh et al. 2005; Sironen et al. 2006) and, by splicing assays, the exons flanked by large introns are found to be the hardest ones to splice consistently (Bell et al. 1998; Fox-Walsh et al. 2005). By contrast, exons flanked by short introns, also associated with high expression levels, tend to be subject to less noisy splicing (Pickrell et al. 2010), and short introns may overcome the poor recognition of alternatively spliced exons (Bell et al. 1998). In addition, exons flanked by long introns tend to be those most commonly lost during evolution (Kandul and Noor 2009), consistent with splice error rates being too high to sustain the exon. To provide a possible explanation, we conjecture that if the flanking intron is large, the splice site is harder to recognize and the possibility for cryptic splice sites contained within the intron would be higher. Thus the true splice sites need to be reinforced by binding of SR proteins to ESEs.

A problem with prior comparative analyses is that they had no non-vertebrate genomes with large introns with which to test a possible coupling between intron size and ESE usage. In chapter II I report on splice-related *cis*-motif usage analysis of the *Ectocarpus* genome, which is a species phylogenetically very distant from vertebrates but, like vertebrates in having abundant large introns. In *Ectocarpus*, patterns of codon and amino acid usage and *k*-mer enrichments in the vicinity of exon boundaries resemble that of

humans. This indicates a deep phylogenetic conservation of exonic splice-related constraints. Moreover, while I identified SR proteins that *Ectocarpus* appears to employ, it would be a helpful follow on study to ask whether certain SR proteins define certain trends through their binding preferences. Likewise it would be good to have experimental confirmation of the ESEs that *Ectocarpus* employs and to see whether the particular SR protein binding motifs are themselves conserved.

My analysis on *Ectocarpus* has left one outstanding mystery, namely why *Ectocarpus* has so many more codons and amino acids showing strong preference avoidance trends (and also so many putative ESEs) than humans. I postulated a low rate of alternative splicing in *Ectocarpus* as a possible factor. At first sight this seems like a good suggestion as *Ectocarpus* is unusual in having multiple large introns and a low rate of alternative splicing, making it all but unique amongst well described genomes. However, when I considered a more phylogenetically extensive and explicit framework (Chapter III) I observed no relationship between alternative splicing rates and *cis*-motif usage. This analysis can certainly be questioned as alternative splicing rate estimation was only possible for a limited sample of the taxa and tends also to be highly dependent on transcriptome depth. The data I employed attempts to avoid the problem of transcriptome depth, but nonetheless I expect that for lesser described species the estimates may well be inaccurate. Nonetheless, this analysis provided no support for the notion that motif usage is a function of alternative splice rates leaving the remarkable enrichment seen in *Ectocarpus* all the more unexpected and unexplained.

Another important finding from *Ectocarpus* genome is that splice optimal and translationally optimal codons are not always mutually exclusive, which is different from what is seen in *Drosophila* (Warnecke and Hurst 2007). One possible reason is that, in *Drosophila*, selection for translational optimality is stronger than that in *Ectocarpus*. This is supported by the observation that, in *Ectocarpus*, the correlation between CAI and expression level is rather weak. Given this, selection to force divergence between translationally optimal and splice optimal codons may also be weak. Another possibility is that the observation in *Drosophila* is an accidental consequence of selection on translational optimality and splice optimality happening to go in opposite directions. We also need to be cautious here as the expression data that I employed is much more limited than that for *Drosophila*. Moreover, repeating the analysis using an array based expression data set I couldn't repeat the finding of a correlation between codon usage and expression level (data not presented). However, this evidence needs to be treated with considerable

caution. On discussing the data with the curator (Simon Dittami) of this data source (A-MEXP-1445 Nimblegen *E.siliculosus* Gene Expression v1: http://www.ebi.ac.uk/arrayexpress/arrays/A-MEXP-1445), I was advised to employ EST-matching method to estimate gene expression, because that microarray is not based on the genome, but on available EST libraries. While the EST data then is the best currently available, strong conclusions need stronger data.

The concentration in Chapter III was not so much the relationship between alternative splicing and ESE usage, but rather more generally the predictors of *cis* motif usage across taxa. As a novel extension to the nearly neutral hypothesis, I suggested that selection can sometimes be stronger when effective population sizes are small (Chapter III). The model proposed that reduced N_e might lead to larger and more introns (following the suggestion of Lynch and Conery), this makes it harder to correctly identify splice sites, which in turn could lead to stronger selection for motifs that inhibit the increase in the degree of error-prone splicing. One ambiguity in our model is the relationship between intron size, ESE enrichment and splice site strength. ESEs are usually found in proximity to weak splice sites. But were the sites weak because of decay (e.g. in low N_e conditions) or weak because ESE enforcement means that it is no longer so crucial to have strong splice sites?

Importantly we find that our measures of $N_e.\mu$ do correlate with intronic dimensions in a phylogenetically explicit framework, whilst the original measures employed by Lynch and Conery do not (as previously pointed out) (Whitney and Garland 2010). It also seems clear that *cis*-motif usage and intronic dimensions covary within and between genomes. As regards the former we find, in the human genome, that there are more ESE-related synonymous sites of exon ends under selection when exons are flanked by larger introns; meanwhile, intronic dimensions and splice site usage predict *cis*-motif usage across species and $N_e.\mu$ predicts intronic dimensions and splice site usage. Furthermore, we find that intra-specifically, exons flanked by large introns both have higher ESE density and greater usage of AGgt, consistent with coevolution between splice site, ESEs and intron size. Therefore, our hypothesis seems to be reasonable as a complement of the nearly neutral hypothesis.

However, some uncertain factors should also be noticed. For example, $N_e.\mu$ is estimated based on polymorphism of orthologous genes, so selection of genes analysed could affect final results. Moreover there might be a systematic issue with all polymorphism based attempts to estimate $N_e.\mu$, this being that the expected correlation between N_e and heterozygozity appears to be much weaker than predicted by the neutral model (which

forms the basis for $N_e.\mu$ estimation), but is consistent with the effects of more common hitchhiking in genomes in large populations (Corbett-Detig, Hartl, and Sackton 2015). Nonetheless, we observe that $N_e.\mu$ robustly predicts intronic dimensions and splice site usage, suggesting that it is perhaps not too poor an estimator.

Although we consider rate of alternative splicing, in our previous study on brown algae *Ectocarpus*, as a factor to explain higher proportion of codons and amino acids showing skews in usage in exon ends (Wu et al. 2013), we do not observe any relationship between *cis*-motif usage and alternative splicing rates in human genome. Also, whether there is a significant difference in ESE density between alternative and constitutive exons is still up in the air (Ke et al. 2011; Cáceres and Hurst 2013). To date, no strong prior evidence has been found that ESE usage is a modulator of alternative splicing and particularly strong purifying selection on ESEs has been excluded as an explanation for the especially low rate of evolution of conserved alternative exons (Parmley, Chamary, and Hurst 2006; Cáceres and Hurst 2013). The current data thus suggest that ESE usage allows determination of splice sites once the "decision" to splice a given exon is made, rather than suggesting that ESEs act as elements to control alternative splicing decisions.

To understand an unexpected result that in the between-species comparison, intron size is by no means the best intron-dimension predictor of *cis*-motif usage (intron density is better than mean intron size, and a combination of size and density is the best predictor), we propose a decoy splice spite model as a potential explanation. This model correctly predicts (or explains) intragenomic and intragenic trends, highlighting the selection on 5' exons being more acute than that on 3' exons. Indeed, this intra-gene bias distribution of selection is consistent with the observation when we investigate occurrence of pathogenic SNPs (Chapter IV).

More generally, when we explore the relationship between determinants of the usage of splice-associated *cis*-motifs and the distribution of human pathogenic SNPs, we find a strong enrichment of pathogenic SNPs at exon ends. From this we infer that these SNPs interfere with splicing. Beyond this as pathogenic SNPs tend to reside in exon with few ESEs, we propose a "fragile" exon model (Chapter IV). Note however, that our inferences here are at arms length – we never actually analyse RNASeq libraries of normal and pathogenic tissues. Detailed analysis in several cases has been performed (Krawczak, Reiss, and Cooper 1992; Nissim-Rafinia and Kerem 2002; Faustino and Cooper 2003; Chamary, Parmley, and Hurst 2006), so a possible connection between splicing and

disease is by no means unexpected. Whether our estimates of the proportion of pathogenic SNPs associated with splicing is correct would require an intensive sampling of splice forms associated with disease causing SNPs, comparing normal and wild type samples to identify the novel splice forms. In addition, to confirm any of these it would be necessary to take an appropriate cell line, remove the wild-type form of the gene, transgenically add the "mis-splicing" form and show deleterious effects consistent with disease pathology. Such experimental analyses are far beyond the domain of this thesis.

My final chapter (Chapter V) is a rather preliminary and exploratory analysis. In this I ask whether tissue specific genes have ESEs that might be especially commonly used in that given tissue. In part the motivation here is biotechnological more than evolutionary. Nonetheless, the finding that some ESEs seem to be especially enriched in genes expressed in certain tissues is worthy of further scrutiny. It would for example, be helpful to see if, when we look at genes expressed in just two tissues we could predict the ESE usage from the analysis of single tissue expressed genes alone. Likewise we can ask whether ESE density as a whole is correlated with tissue specificity. A particularly elegant way of confirming my preliminary observation would be to consider tissue specific exons and ask if these too use the "tissue specific ESEs".

References

Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, Kappes J, Collawn JF, Bebok Z. 2010. A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J Biol Chem* 285:28741-28748.

Bell MV, Cowper AE, Lefranc MP, Bell JI, Screaton GR. 1998. Influence of intron length on alternative splicing of CD44. *Mol Cell Biol* 18:5930-5941.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hébuterne X, et al. 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* 43:242-245.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol* 62:89-98.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6:R75.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7:98-108.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol* 13:e1002112.

Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev* 17:419-437.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* 102:16176-16181.

Gartner JJ, Parker SC, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, et al. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A* 110:13481-13486.

Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J Mol Evol* 63:174-182.
Kandul NP, Noor MA. 2009. Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC Genet* 10:67.
Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 21:1360-1374.
Klinz FJ, Gallwitz D. 1985. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 13:3791-3804.
Krawczak M, Reiss J, Cooper DN. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 90:41-54.
Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in D. melanogaster. *PLoS Genet* 9:e1003527.
Luehrsen KR, Walbot V. 1992. Insertion of non-intron sequence into maize introns interferes with splicing. *Nucleic acids research* 20:5181-5187.
Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314:1930-1933.
Nissim-Rafinia M, Kerem B. 2002. Splicing regulation as a potential genetic modifier. *Trends Genet* 18:123-127.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* 23:301-309.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* 5:e14.
Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* 6:e1001236.
Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J. 2006. An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A* 103:5006-5011.
Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Molecular Biology and Evolution* 24:2755-2762.
Warnecke T, Weber CC, Hurst LD. 2009. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Trans* 37:756-761.
Whitney KD, Garland T, Jr. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet* 6.
Wu X, Tronholm A, Cáceres EF, Tovar-Corona JM, Chen L, Urrutia AO, Hurst LD. 2013. Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga Ectocarpus are common and resemble those seen in humans. *Genome Biol Evol* 5:1731-1745.